Chapter 18: National Assessment of Educational Progress (NAEP)

1. OVERVIEW

BIENNIAL SURVEY OF A SAMPLE OF ELEMENTARY/ SECONDARY STUDENTS

Two assessments:

- Main NAEP
- Trend NAEP

Three foci:

- Main National NAEP
- Main State NAEP
- Trial Urban District NAEP

Four component surveys:

- Elementary and Secondary School Students Survey
- School Characteristics and Policies Survey
- Teacher Survey
- SD/ELL Survey/ Excluded Student Survey

The National Assessment of Educational Progress (NAEP) is mandated by Congress to assess the educational achievement of U.S. students and monitor changes in those achievements. As the only nationally representative and continuing assessment of what America's students know and can do in nine subject areas, NAEP serves as the "Nation's Report Card." The *main national* NAEP regularly assesses the achievement of students in grades 4, 8, and 12 at the national level. The *main state* NAEP assessed students at both grades 4 and 8 in at least one subject in 1990, 1992, 1994, 1996, 1998, 2000, 2002, and 2003. Since 2003, the main state NAEP has assessed students in at least two subjects, reading and mathematics, every 2 years at grades 4 and 8. The NAEP *Trial Urban District Assessment (TUDA)* assessed performance in selected large urban districts in 2002 in reading and writing at grades 4 and 8, and continued in 2003, 2005, 2007 and 2009 with reading and mathematics assessments at grades 4 and 8, and alternately science or writing.The *trend* NAEP tracks national long-term trends since the 1970s in mathematics and reading at ages 9, 13, and 17, and is given every 4 years. The national assessments were first implemented in 1969 and were conducted on an annual or biennial basis through 1995, and annually since 1996. The state assessments have been administered biennially since 1990.

In 1988, Congress established the National Assessment Governing Board (NAGB) to provide policy guidance for the execution of NAEP. The 26-member Governing Board is an independent, bipartisan group whose members include governors, state legislators, local and state school officials, educators, business representatives, and members of the general public. Its responsibilities include: select subject areas to be assessed; set appropriate student achievement levels; develop assessment objectives and test specifications; design the assessment methodology; and produce standards and procedures for interstate, regional, and national comparisons. NAEP is administered by the National Center for Education Statistics (NCES).

Purpose

To (1) monitor continuously the knowledge, skills, and performance of the nation's children and youth; and (2) provide objective data about student performance at the national, the regional, the state level (since 1990), and the district level (since 2002).

Components

NAEP comprises two unique assessments: *main and trend*; and there are three foci in the *main* assessment: *main national*, *main state* and *trial urban district*. Each of these assessments consists of four components: Elementary and Secondary School Students Survey; School Characteristics and Policies Survey; Teacher Survey; and

Students with Disabilities or English language learners(SD/ELL) Survey (for the *main national* NAEP) or Excluded Student Survey (for the *trend* NAEP).

In 1985, the Young Adult Literacy Study was also conducted nationally as part of NAEP, under a grant to the Educational Testing Service and Response Analysis Corporation; this study assessed the literacy skills of 21- to 25-year-olds. In addition, a High School Transcript Study (HSTS, see chapter 29) and a National Indian Education Study (NIES) are periodically conducted as components of NAEP.

Since 1996, the main national and state assessments have included accommodations for students with special needs.

National-level assessment. The *main national* NAEP and *trend* NAEP are both designed to report information for the nation and specific geographic regions of the country (Northeast, Southeast, Central, and West). However, these two assessments use separate samples of students from public and nonpublic schools: grade samples for the main national NAEP (grades 4, 8, and 12), and age/grade samples for trend NAEP (age 9/grade 4; age 13/grade 8; age 17/grade 11). The test instruments for the two assessments are based on different frameworks; the student and teacher background questionnaires vary; and the results for the two assessments are reported separately. (See "Elementary and Secondary School Students Survey" below for the subject areas assessed.)

The assessments in the *main national* NAEP follow the curriculum frameworks developed by NAGB and use the latest advances in assessment methodology. The test instruments are flexible so they can be adapted to changes in curricular and educational approaches. Recent assessment instruments for the main NAEP have been kept stable for short periods of time, allowing short-term trends to be reported from 1990 through 2009, except for the mathematics assessment for grade 12. In 2005, and 2009, NAGB introduced changes in the NAEP mathematics framework for grade 12 in both the assessment content and administration procedures.

To reliably measure change over longer periods of time, the *trend* NAEP must be used. For long-term trends, past procedures must be precisely replicated with each new assessment, and the survey instruments do not evolve with changes in curricula or educational practices. The instruments used today for the trend NAEP are relatively identical to those developed in the 1970s. Trend NAEP allows measurement of trends since 1971 in reading and 1973 in mathematics.

State-level assessments. The *main state* NAEP was implemented in 1990 on a trial basis and has been conducted biennially since that time. Participation of the states was completely voluntary until 2003. The reauthorization of the Elementary and Secondary Education Act, also referred to as the "No Child Left Behind Act," requires states that receive Title I funding to participate in state NAEP assessments in reading and mathematics at grades 4 and 8 every 2 years. State participation in other state NAEP subjects (i.e., science and writing) remains voluntary. Separate representative samples of students are selected for each jurisdiction to provide that jurisdiction with reliable state-level data concerning the achievement of its students. The state assessment included nonpublic schools in 1994, 1996, and 1998. This practice ended because of low participation rates. (See below for the subject areas assessed.)

The Trial Urban District Assessment. The *Trial Urban District Assessment (TUDA)* began assessing performance in selected large urban districts in 2002 in reading and writing; it continued in 2003 with reading and mathematics; in 2005 with reading, mathematics, and science; in 2007 with reading, mathematics and writing, and in 2009 with reading, mathematics and science. The program retains its trial status. The first TUDA occurred in reading and writing in 2002 for five urban districts. In 2003, nine districts were assessed in mathematics and reading. In 2005 and 2007, ten urban school districts participated in TUDA. The results for these districts are for public school students only. Results for District of Columbia public school students, normally included with NAEP's state assessment results, are also reported in TUDA in 2005 and 2007 in reading and mathematics. (Due to an insufficient sample size, the District of Columbia did not participate in the science assessment in 2005 and 2009 and the writing assessment in 2007.) Beginning in 2009, the TUDA results include only those charter schools that the district is accountable for.) Results for these districts are also compared with results for public school students in large central cities and the nation.

Elementary and Secondary School Students Survey. The primary data collected by NAEP relate to student performance and educational experience as reported by students. Major assessment areas include: reading, writing, mathematics, science, civics, U.S. history, geography, economics, and the arts.

Subjects assessed in the main national NAEP. In 1988, the *main national* NAEP assessed student performance in reading, writing, civics, and U.S. history, and conducted small assessments in geography and document literacy. In 1990, it assessed mathematics, reading, and science; in 1992, reading, mathematics, and writing; in 1994, reading, U.S. history, and world geography; and in 1996, science and mathematics. A probe of student performance in the arts at grade 8 was conducted in 1997. Reading, writing, and civics were assessed in 1998. (*Trend* NAEP was conducted in 1999.) In 2000, the main national NAEP assessed mathematics and science (and, for 4th-graders only, reading). In 2001, history and geography were assessed; in 2002, reading and writing. In 2003, the assessments were in reading and mathematics for 4th- and 8th-graders. In 2004, the main national NAEP assessed foreign language for 12th grade. In 2005, the assessments were in reading, mathematics, and science, and in 2006, in U.S. history and civics (and, for 12th-graders only, in economics). In 2007, reading and mathematics were assessed at grades 4 and 8, and writing at grades 8 and 12. In 2008, the arts were assessed at grade 8. In 2009, reading, mathematics and science were assessed at grade 4, 8 and 12. In 2010, U.S. history, civics and geography were assessed at grade 4, 8 and 12.

Subjects assessed in trend NAEP. The subjects assessed in *trend* NAEP are mathematics and reading (and, until 1999, writing; and, until 2004, science). The biennial assessments from 1988 through 1996 covered all subjects. Since 2004, the trend assessments have been scheduled to be administered in mathematics and reading every 4 years. The latest trend assessment was conducted in 2008, and the report was released in the spring of 2009.

Subjects assessed in the main state NAEP. Data representative of states were collected for the first time in the 1990 trial state assessment, when 8th-grade students were assessed in mathematics. In 1992, state-level data were collected in 4th-grade reading and mathematics, and in 8th-grade mathematics. In 1994, 4th-grade reading was assessed. In 1996, 4th-grade mathematics and 8th-grade mathematics and science were assessed. The 1998 NAEP collected state-level data in reading at grades 4 and 8, and writing at grade 8. The 2000 NAEP assessments covered mathematics and science, the 2002 assessments covered reading and writing, the 2003 assessments covered reading and mathematics, and the 2005 assessment covered reading, mathematics, and science. The 2007 state assessment covered reading and mathematics (and, for grade 8 only, writing). The 2009 state assessment covered reading, mathematics and science at grades 4 and 8, and reading and mathematics at grade 12.

Subjects assessed at TUDA. Data representative of urban districts were collected for the first time on a trial basis in selected large urban districts in 2002 with reading and writing assessments. In 2003, district-level data were collected in 4th- and 8th-grade reading and mathematics. In 2005, 4th- and 8th-grade reading, mathematics and science were assessed. In 2007, 4th- and 8th-grade reading, mathematics and writing were assessed. TUDA retains its trial status, and is scheduled for 2009 .

Student background questions. The student survey also asks questions about the student's background, as well as questions related to the subject area and the student's motivation in completing the assessment. Student background questions gather information about race/ethnicity, school attendance, academic expectations, and factors believed to influence academic performance, such as homework habits, the language spoken in the home, and the quantity of reading materials in the home. Some of these questions document changes that occur over time: these questions remain unchanged over assessment years.

Student subject-area questions. These questions gather three categories of information: time spent studying the subject, instructional experiences in the subject, and perceptions about the subject. Because these questions are specific to each subject area, they can probe in some detail the use of specialized resources (such as the use of calculators in mathematics classes).

Students are also asked how often they have been asked to write long answers to questions on tests or assignments that involve the tested subject. Before 2004, students were also asked how many questions they thought they answered correctly, how difficult they found the assessment, how hard they tried on this test compared to how hard they had tried on other tests or assignments they had taken that year in school, and how important it was to them to do well on this test.

School Characteristics and Policies Survey. This survey collects supplemental data about school characteristics and school policies that can be used analytically to provide context for student performance issues. Data are collected on enrollment, absenteeism, dropout rates, curricula,

testing practices, length of school day and year, school administrative practices, school conditions and facilities, size and composition of teaching staff, tracking policies, schoolwide programs and problems, availability of resources, policies for parental involvement, special services, and community services.

Teacher Questionnaire. This study collects supplemental data from teachers whose students are respondents to the assessment surveys. The first part of the teacher questionnaire tends to cover background and general training, and includes items concerning years of teaching experience, certifications, degrees, major and minor fields of study, course work in education, course work in specific subject areas, the amount of in-service training, the extent of control over instructional issues, and the availability of resources for the classroom. Subsequent parts of the teacher questionnaire tend to cover training in the subject area, classroom instructional information, and teacher exposure to issues related to the subject and the teaching of the subject. They also ask about pre- and in-service training, the ability level of the students in the class, the length of homework assignments, use of particular resources, and how students are assigned to particular classes.

***SD/ELL Survey*.** This survey is completed in the *main* NAEP assessments (and the *trend* NAEP since 2004) by teachers of students who are selected to participate in NAEP but who are classified as either having disabilities (SD) or English language learners (ELL). Information is collected on the background and characteristics of each SD/ELL student and the reason for the SD/ELL classification, as well as on whether these students receive accommodations in district or statewide tests. For SD students, questions ask about the student's functional grade levels and special education programs. For ELL students, questions ask about the student's native language, time spent in special language programs, and level of English language proficiency. This survey is used to determine whether the student should take the NAEP assessment. If any doubt exists about a student's ability to participate in the assessment, the student is included. Beginning with the 1996 assessments, NAEP has allowed accommodations for both SD and ELL students.

Excluded Student Survey. This survey is completed in trend NAEP for students who are sampled for the assessment, but who are excluded by the school from participating in it. Following exclusion criteria used in previous trend assessments, a school can exclude students with limited English-speaking ability, students who are educable mentally retarded, and students who are functionally disabled—if the school judges that these students are unable to "participate meaningfully" in the assessment. This survey is only completed for those students who are actually excluded from the assessment (whereas the SD/ELL Survey in the main assessment is also completed for participating students who are SD or ELL students—see above).

High School Transcript Study. Transcript studies have been conducted in 1987, 1990, 1994, 1998, 2000, 2005, and 2009. The studies collect information on current course offerings and course-taking patterns in the nation's schools. Transcript data can be used to show course-taking patterns across years that may be associated with proficiency in subjects assessed by NAEP. Transcripts are collected from grade 12 students in selected schools in the NAEP sample. (For more information on the High School Transcript Studies, see chapter 29.)

National Indian Education Study. The National Indian Education Study (NIES) is a two-part study designed to describe the condition of education for American Indian and Alaska Native (AI/AN) students in the United States. The study is conducted by NCES on behalf of the U.S. Department of Education, Office of Indian Education. NIES is authorized under Executive Order 13336, "American Indian and Alaska Native Education", which was signed in 2004 to improve education efforts for AI/AN students nationwide.

Part I of NIES is conducted through NAEP and provides in-depth information on the academic performance of 4th– and 8th–grade AI/AN students in reading and mathematics. Part II of NIES is a survey that describes the educational experiences of the 4th– and 8th–grade AI/AN students who participated in the NAEP assessments. The survey focuses on the integration of native language and culture into school and classroom activities. Part II collects information through questionnaires for students, teachers, and principals.

Oral Reading Study. In 2002, NAEP conducted a special study on oral reading. The NAEP 2002 Oral Reading Study looked at how well the nation's 4^{th}-graders can read aloud a grade-appropriate story. NAEP assessed a random sample of 4^{th}-grade students selected for the NAEP 2002 reading and writing assessments. The assessment provided information about a student's fluency in reading aloud and examined the relationship between oral reading

accuracy, rate (or speed), fluency, and reading comprehension.

Technology-Based Assessment (TBA) Project. TBA was a NAEP project in 2000 to 2003.TBA was designed with five components—three empirical studies (Mathematics Online, Writing Online, and Problem Solving in Technology-Rich Environment); a conceptual paper (Computerized Adaptive Testing); and an online school and teacher questionnaire segment. The three empirical studies were the primary focus of the TBA Project and are discussed below.

The primary goals of the Mathematics Online (MOL) study were to understand how computer delivery affects the measurement of NAEP math skills, to gain insight into the operational and logistical mechanics of computer-delivered assessments, and to evaluate the ability of 4th- and 8th-graders to deal with mathematics assessments delivered on computer. At grade 8, an additional goal was to investigate the technical feasibility of generating alternate versions of multiple-choice and constructed-response items using "on-the-fly" technology. MOL was field tested in 2002.

The Writing Online (WOL) study was intended to help NAEP learn how computer delivery affects the measurement of NAEP performance-based writing skills, to gain insights into the operational and logistical mechanics of computer-delivered writing assessments, and to evaluate the ability of 8th-graders to deal with writing assessments delivered on computer. WOL was field tested in 2002.

The Problem Solving in Technology-Rich Environments (TRE) study was designed to develop an example set of modules to assess problem solving using technology. These example modules use the computer to present multimedia tasks that cannot be delivered through conventional paper-and-pencil assessments, but which tap important emerging skills. TRE was field tested in 2003.

Charter School Pilot Study. NAEP conducted a pilot study of America's charter schools and their students as part of the 2003 NAEP assessments in reading and mathematics at the 4th-grade level. Charter schools are public schools of choice. They serve as alternatives to the regular public schools to which students are assigned. While there are many similarities between charter schools and other public schools, they do differ in some important ways, including the makeup of the student population and their location.

Student Achievement in Private Schools. To better understand the performance of students in private schools, NAEP performed two studies and has released a two-part series of reports. In the first report *Student Achievement in Private Schools: Results from NAEP 2000–2005 (*Perie, Vanneman, & Goldstein 2005), the results of the 2000, 2002, 2003, and 2005 assessments for all private schools and for the largest private school categories—Catholic, Lutheran, and conservative Christian—were compared with the results for public schools (where applicable). This report focused on important demographic differences between students nationwide in private and public schools. The goal of the second report (*Comparing Private Schools and Public Schools Using Hierarchical Linear Modeling* [Braun, Jenkins, & Grigg 2006]) was to examine differences in mean NAEP reading and mathematics scores in 2003 between public and private schools when selected characteristics of students and/or schools were taken into account. Hierarchical linear models were employed to carry out the desired adjustments.

Periodicity

Annual from 1969 to 1979, biennial in even-numbered years from 1980 to 1996, after which it was annual. A probe of 8th-graders in the arts was conducted in 1997 and again in 2008. State-level assessments, initiated in 1990, follow the same schedule as the main national assessments. Prior to 1990, NAEP was required to assess reading, mathematics, and writing at least once every 5 years. The previous legislation required assessments in reading and mathematics at least every 2 years, in science and writing at least every 4 years, and in history or geography and other subjects selected by the NAGB at least every 6 years.

The No Child Left Behind Act requires NAEP to conduct national and state assessments at least once every 2 years in reading and mathematics in grades 4 and 8. In addition, NAEP has conducted a national assessment in reading and mathematics in grade 12 every 4 years starting since 2005. TUDA began assessing performance in selected large urban districts in 2002 with reading and writing assessments and continued in 2003, 2005, 2007 and 2009 with reading and mathematics assessments. TUDA is scheduled for 2011 as well. The program retains its trial status. Finally, to the extent that time and money allow, NAEP will be conducted in grades 4, 8, and 12 at regularly scheduled intervals in additional subjects including writing, science, history, geography, civics, economics, foreign languages, and the arts.

NIES was conducted for the first time in 2005 as a part of NAEP, in accordance with Title VII, Part A of the Elementary and Secondary Education Act, 2001. The second NIES data collection took place in 2007, and

the third collection took place in 2009: NCES is planning to conduct NIES again in 2011.

2. USES OF DATA

NAEP is the only ongoing, comparable, and representative assessment of what American students know and can do in nine subject areas. Policymakers are keenly interested in NAEP results because they address national outcomes of education, specifically, the level of educational achievement. In addition, state-level and urban district-level data, available for many states since 1990 and for selected large urban districts since 2002, allow both state-to-state and district-to-district comparisons, and comparisons of individual states with the nation as a whole (as well as comparisons of urban districts with large central cities and the nation).

During NAEP's history, a number of reports across various subject areas have provided a wealth of information on students' academic performance, learning strategies, and classroom experiences. Together with the performance results, the basic descriptive information collected about students, teachers, administrators, and communities can be used to address the following educational policy issues:

- *Instructional practices.* What instructional methods are being used?
- *Students-at-risk.* How many students appear to be at-risk in terms of achievement, and what are their characteristics? What gaps exist between at-risk categories of students and others?
- *Teacher workforce.* What are the characteristics of teachers of various subjects?
- *Education reform.* What policy changes are being made by our nation's schools?

However, *users should be cautious in their interpretation of NAEP results. While NAEP scales make it possible to examine relationships between students' performance and various background factors, the relationship that exists between achievement and another variable does not reveal its underlying cause, which may be influenced by a number of other variables.* NAEP results are most useful when they are considered in combination with other knowledge about the student population and the education system, such as trends in instruction, changes in the school-age population, and societal demands and expectations.

NAEP materials such as frameworks and released questions also have many uses in the educational community. Frameworks present and explain what experts in a particular subject area consider important. Several states have used NAEP frameworks to revise their curricula. After most assessments, NCES publicly releases nearly one-third of the questions. Released constructed-response questions and their corresponding scoring guides have served as models of innovative assessment practices in the classroom.

3. KEY CONCEPTS

The achievement levels for NAEP assessments are defined below. For subject-specific definitions of achievement levels and additional terms, refer to NAEP technical reports, "report card" reports, and other publications.

Achievement levels. Starting with the 1990 NAEP, NAGB developed achievement levels for each subject at each grade level to measure how well students' actual achievement matches the achievement desired of them. The three levels are as follows:

- *Basic.* Partial mastery of the prerequisite knowledge and skills that are fundamental for proficient work at each grade.
- *Proficient.* Solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
- *Advanced.* This level signifies superior performance, and is attained by only a very small percentage of students (3–6 percent) at any of the three grade levels assessed.

4. SURVEY DESIGN

Target Population

Students enrolled in public and nonpublic schools in the 50 states and the District of Columbia who are deemed assessable by their school and classified in

defined grade/ age groups—grades 4, 8, and 12 for the *main national* assessments and ages 9, 13, and 17 for the *trend* assessments in science, mathematics, and reading. Grades 4 and/or 8 are usually assessed in the *state* assessments and *TUDA*; the number of grades assessed has varied in the past, depending on the availability of funding (although testing for 4th- and 8th-graders in reading and mathematics every 2 years is now required for states that receive Title I funds). Only public schools were included in the *state* NAEP prior to 1994 and after 1998. Only public schools are included in *TUDA*.

Sample Design

For the national assessments, probability samples of schools and students are selected to represent the diverse student population in the United States. The numbers of schools and students vary from cycle to cycle, depending on the number of subjects and items to be assessed. A national sample will have sufficient schools and students to yield data for public schools and each of the four NAEP regions of the country, as well as sex, race, degree of urbanization of school location, parent education, and participation in the National School Lunch Program. A separate grade 12 sample of schools is also selected to produce national and regional estimates, as state NAEP does not yet include grade 12 (a pilot study of grade 12 state NAEP was conducted in 2009). A national sample of nonpublic (private) schools is also selected for grades 4, 8, and 12. This sample is designed to produce national and regional estimates of student performance for private schools.

In the state assessment, a sample of schools and students is selected to represent a participating state. In a state, on average 2,500 students in approximately 100 public schools are selected per grade, per subject assessed. The selection of schools is random within classes of schools with similar characteristics; however, some schools or groups of schools (districts) can be selected for each assessment cycle if they are unique in the state. For instance, a particular district may be selected more often if it is located in the state's only major metropolitan area or has the majority of the state's Black, Hispanic, or other race/ethnicity population. Additionally, even if a state decides not to participate at the state level, schools in that state identified for the national sample will be asked to participate.

Typically, 30 students per subject per grade are selected randomly in each school. Some of the students who are randomly selected are classified as SD or ELL. NAEP's goal is to assess all students in the sample, and this is done if at all possible.

NAEP's multistage sampling process involves the following steps:

- selection of schools (public and nonpublic) within strata and
- selection of students within the selected schools.

Selection of schools. In this stage of sampling, public schools in each state (—including Bureau of Indian Education [BIE] schools and Department of Defense Education Activity [DoDEA] schools)—and nonpublic schools in each state (including Catholic schools) are listed according to the grades associated with the three age classes: age class 9 refers to age 9 or grade 4 in the trend NAEP (or grade 4 in the main NAEP); age class 13 refers to age 13 or grade 8 in the trend NAEP (or grade 8 in the main NAEP); age class 17 refers to age 17 or grade 11 in the trend NAEP (or grade 12 in the main NAEP).

The school lists are obtained from two sources. Regular public, BIE, and DoDEA schools are obtained from the school list maintained by Common Core of Data. (See chapter 2). Catholic and other nonpublic schools are obtained from the NCES Private School Universe Survey (PSS). (See chapter 3.) To ensure that the state samples provide an accurate representation, public schools are stratified by urbanization, enrollment of Black, Hispanic, or other race/ethnicity students, and median house-hold income. Nonpublic schools are stratified by type of control (e.g., parochial, nonreligious), urban status, and enrollment per grade. Once the stratification is completed, the schools within each state are assigned a probability of selection that is proportional to the number of students per grade in each school.

Prior to 2005, DoDEA overseas and domestic schools were reported separately. In the 2005 assessments, all DoDEA schools, both domestic and overseas, were combined into one jurisdiction. In addition, the definition of the national sample changed in 2005; it now includes all of the overseas DoDEA schools.

The manner of sampling schools for the long-term trend assessments is very similar to that used for the main assessments. The primary difference is that in long-term trend nonpublic schools and schools with high enrollment of Black, Hispanic, or other race/ethnicity students are not oversampled. Schools

are not selected for both main and long-term trend assessments at the same age/grade. The long-term trend assessments use a nationally representative sample and do not report results by state.

Selection of students. This stage of sampling involves random selection of national samples representing the entire population of U.S. students in grades 4, 8, and 12 for the main assessment and the entire population of students at ages 9, 13, and 17 for the long-term trend assessment. Typically, 30 students per subject per grade are selected randomly in each school. Some of the students who are randomly selected are classified as SD or ELL. A small number of students selected for participation are excluded because of limited English proficiency or severe disability.

To facilitate the sampling of students, a consolidated list is prepared for each school of all age-eligible students (long-term trend assessments) or all grade-eligible students (main assessments) for the age class for which the school is selected. A systematic selection of eligible students is made from this list—unless all students are to be assessed—to provide the target sample size.

For each age class (separately for long-term trend and main samples), maxima are established as to the number of students who are to be selected for a given school. In those schools that, according to information in the sampling frame, have fewer eligible students than the established maxima, each eligible student enrolled at the school is selected in the sample. In other schools, a sample of students is drawn. The maximum sample sizes are established in terms of the number of grade-eligible students for the main samples, and in terms of the number of students in each age class for the trend samples.

Excluded students Some students are excluded from the student sample because they are deemed unassessable by school authorities. The exclusion criteria for the main samples differ somewhat from those used for the long-term trend samples. In order to identify students who should be excluded from the main assessments, school staff members are asked to identify those SD or ELL students who do not meet the NAEP inclusion criteria. School personnel are asked to complete an SD/ELL questionnaire for all SD and ELL students selected into the NAEP sample, whether they participate in the assessment or not. Prior to 2004, for the long-term trend assessments, excluded students were identified for each age class, and an Excluded Student Survey was completed for each excluded student. Beginning in 2004, both trend and main NAEP assessments use identical procedures.

For the special study of Students with Disabilities or Limited English Proficient (SD/LEP) inclusion in the 1996 main assessment, oversampling procedures were applied to SD/LEP students at all three grades in sample types 2 (accommodations not allowed) and 3 (accommodations allowed) for mathematics and in sample type 3 for science. (Sample type denotes whether or not a session may allow such accommodations.)

Main national NAEP sample sizes. Not all subject areas are assessed in every assessment year. In 2009, the main national NAEP assessed students in reading, mathematics and science at grades 4, 8 and 12. For the main national NAEP, a nationally representative sample of more than 350,000 students at grades 4, 8, and 12 participated in these assessments. The main national math assessment sampled 168,800 4^{th} grade students, 161,700 8^{th} grade students, and 48,900 12^{th} grade students; the reading assessment sampled 178,800 4^{th} grade students, 160,900 8^{th} grade students, and 51,700 12^{th} grade students. The science assessment sampled 156,500 4^{th} grade students, 151,100 8^{th} grade students, and 11,100 12^{th} grade student. The mathematics, reading, and science assessments were conducted in the same 9,600 4^{th} grade schools, 7,110 8^{th} grade schools, and 1,680 12^{th} grade schools.

TUDA sample sizes. In 2009, eighteen urban districts (including District of Columbia) participated in TUDA in math and reading and 17 urban districts participated in TUDA in science. The sample design for TUDA districts provides for oversampling. For the five largest TUDA districts—New York City, Los Angeles, Chicago, Miami and Houston—the target student sample sizes are three-quarters the normal size of the state sample. For the other twelve districts (Atlanta, Austin, Baltimore City, Boston, Charlotte, Cleveland, Detroit, Fresno, Jefferson County, KY, Milwaukee, Philadelphia, and San Diego), the target student sample sizes are half the normal size of the state sample. The larger samples allow reliable reporting about subgroups in these districts.

Students in the TUDA samples are considered part of the state and national samples. For example, the data for students tested in the Chicago sample will be used to report results for Chicago, but will also contribute to Illinois' estimates (and, with appropriate weights, to national estimates). Chicago has approximately 20 percent of the students in Illinois; therefore Chicago

will contribute 20 percent, and the rest of the state will contribute 80 percent, to Illinois' results.

Long-term trend NAEP sample sizes. The long-term trend assessment tested the same four subjects across years through 1999, using relatively small national samples. Samples of students were selected by age (9, 13, and 17) for mathematics, science, and reading, and by grade (4, 8, and 11) for writing. Students within schools were randomly assigned to either mathematics/science or reading/writing assessment sessions subsequent to their selection for participation in the assessments. In 2004, science and writing were removed from the trend assessments; the trend assessments are now scheduled to be administered in mathematics and reading every 4 years (but not in the same years as the main assessments). In 2004, approximately 24,100 students took the modified[1] reading assessment, while about 14,000 took the bridge[2] reading assessment. In 2004, approximately 22,400 students took the modified mathematics assessment, while about 14,700 took the bridge mathematics assessment. The latest trend assessment was conducted in 2008, with approximately 26,600 students assessed in reading, and 26,700 students assessed in mathematics.

NIES Part II sample sizes. The NIES Part II sample is designed to produce information representative of the target population of all fourth- and eighth-grade AI/AN students in the United States. In 2005, the sample included about 5,600 eligible students at approximately 550 schools located throughout the United States. The sample consisted of approximately 84 percent public, 4 percent private, and 12 percent BIE schools (unweighted). In 2007, the NIES Part II sample included about 12,900 AI/AN students at approximately 1,900 schools at grade 4 and 14,600 AI/AN students at 2,000 schools at grade 8 located throughout the United States. The sample consisted of approximately 94 percent public, 1 percent private, and 5 to 6 percent BIE schools at grades 4 and 8 (as well as a small number of DoDEA schools). All BIE schools were part of the sample. In 2009, the NIES Part II sample consisted of about 12,300 grade 4 students in approximately 2,300 schools and approximately 10,400 students in grade 8 at about 1,900 schools.

Assessment Design

Since 1988, NAGB has selected the subjects for the main NAEP assessments. NAGB also oversees the creation of the frameworks that underlie the assessments and the specifications that guide the development of the assessment instruments.

Development of framework and questions. NAGB uses an organizing framework for each subject to specify the content that will be assessed. This framework is the blueprint that guides the development of the assessment instrument. The framework for each subject area is determined with input from teachers, curriculum specialists, subject-matter specialists, school administrators, parents, and members of the general public.

Unlike earlier multiple-choice instruments, current instruments dedicate a majority of testing time to constructed-response questions that require students to compose written answers. Constructed-response questions provide a separate means of assessing ability that taps recall, not recognition.

The questions and tasks in an assessment are based on the subject-specific frameworks. They are developed by teachers, subject-matter specialists, and testing experts under the direction of NCES and its contractors. For each subject-area assessment, a national committee of experts provides guidance and reviews the questions to ensure that they meet the framework specifications. For each state-level assessment, state curriculum and testing directors review the questions that will be included in the NAEP state component.

Matrix sampling. Several hundred questions are typically needed to reliably test the many specifications of the complex frameworks that guide NAEP assessments. However, administering the entire collection of cognitive questions to each student would be far too time consuming to be practical. Matrix sampling allows the assessment of an entire subject area within a reasonable amount of testing time, in most cases 50 minutes. By this method, different portions from the entire pool of cognitive questions are printed in separate booklets and administered to different but equivalent samples of students.

The type of matrix sampling used by NAEP is called focused, balanced incomplete block (BIB) spiraling. The NAEP BIB design varies according to subject area.

Data Collection and Processing

Since 1983, NCES has conducted NAEP through a series of contracts, grants, and cooperative agreements with the Educational Testing Service

[1] The modified assessment included new items and features, representing the new design.
[2] The bridge assessment replicates the assessment given in the previous assessment year.

(ETS) and other contractors. ETS is directly responsible for developing the assessment instruments, analyzing the data, and reporting the results. Westat selects the school and student samples, trains assessment administrators, and manages field operations (including assessment administration and data collection activities). NCS Pearson is responsible for printing and distributing the assessment materials and for scanning and scoring students' responses.

Reference dates. Data for the main national NAEP and main state NAEP are collected from the last week in January through the first week in March. Data for the long-term trend NAEP are collected during the fall for age 13; during the winter of the same school year for age 9; and during the spring for age 17.

Data collection. Before 2002, NCES had relied heavily on school administrators for the conduct of main state NAEP assessments. Beginning with the 2002 assessments, however, NAEP contractor staff has conducted all NAEP assessment sessions. Obtaining the cooperation of the selected schools requires substantial time and energy, involving a series of mailings that includes letters to the chief state school officers and district superintendents to notify the sampled schools of their selection; additional mailings of informational materials; and introductory in-person meetings where procedures are explained.

The questionnaires for the School Characteristics and Policies Survey, the Teacher Survey, and the SD/ELL Survey are sent to schools ahead of the assessment date so that they can be collected when the assessment is administered. Questionnaires not ready at this time are retrieved later, either through a return visit by NAEP personnel or through the mail.

NCS Pearson produces the materials needed for NAEP assessments. NCS Pearson prints identifying barcodes and numbers for the booklets and questionnaires, pre-assigns the booklets to testing sessions, and prints the booklet numbers on the administration schedule. These activities improve the accuracy of data collection and assist with the BIB spiraled distribution process.

Assessment exercises are administered either to individuals or to small groups of students by specially trained field personnel. For all three ages in the long-term trend NAEP, the mathematics questions administered using a paced audiotape before 2004. Since 2004, the long-term trend assessments have been administered through test booklets read by the students.

For the long-term trend assessments, Westat hires and trains approximately 85 field staff to collect the data. For the 2009 main national and state assessments, Westat hired and trained about 7,000 field staff to conduct the assessments.

After each session, Westat staff interview the assessment administrators to receive their comments and recommendations. As a final quality control step, a debriefing meeting is held with the state supervisors to receive feedback that will help improve procedures, documentation, and training for future assessments.

For NIES Part II, NCES data collection contractor staff visit the schools to administer survey questionnaires. Students complete the questionnaires in group settings proctored by study representatives. In order to decrease the possibility that survey responses might be adversely affected by students' reading levels, the questions are read aloud to all grade 4 students and to grade 8 students who school staff think might need assistance. In addition, the study representatives are available to answer any questions that students have as they work on the questionnaires.

In 2005, survey materials were mailed to about 20 percent (unweighted) of the NIES Part II schools (primarily schools that were remotely located and had only a few AI/AN students), and the schools were asked to administer the questionnaires and return them by mail. Detailed instructions were provided for identifying teachers and students to be surveyed, administering the student questionnaires, responding to questions from students, and labeling and returning survey materials. Although the mail mode was used at about 20 percent (unweighted) of the sampled schools, these schools generally had only one or two sampled students. Thus, only about 2 percent of the sampled students were at mail-mode schools. The mail-mode data collection procedure was discontinued after the 2005 administration of NIES Part II.

Data processing. NCS Pearson handles all receipt control, data preparation and processing, scanning, and scoring activities for NAEP. Using an optical scanning machine, NCS Pearson staff scans the multiple-choice selections, the handwritten student responses, and other data provided by students, teachers, and administrators. An intelligent data entry system is used for resolution of the scanned data, the entry of documents rejected by the scanning machine, and the entry of information from the

questionnaires. An image-based scoring system introduced in 1994 virtually eliminates paper handling during the scoring process. This system also permits online monitoring of scoring reliability and creation of recalibration sets.

ETS and NCS Pearson develop focused, explicit scoring guides with defined criteria that match the criteria emphasized in the assessment frameworks. The scoring guides are reviewed by subject-area and measurement specialists, the instrument development committees, NCES, and NAGB to ensure consistency with both question wording and assessment framework criteria. Training materials for scorers include examples of student responses from the actual assessment for each performance level specified in the guides. These exemplars help scorers interpret the scoring guides consistently, thereby ensuring the accurate and reliable scoring of diverse responses.

The image-based scoring system allows scorers to assess and score student responses online. This is accomplished by first scanning the student response booklets, digitizing the constructed responses, and storing the images for presentation on a large computer monitor. The range of possible scores for an item also appears on the display; scorers click on the appropriate button for quick and accurate scoring. The image-based scoring system facilitates the training and scoring process by electronically distributing responses to the appropriate scorers and by allowing ETS and NCS Pearson staff to monitor scorer activities consistently, identify problems as they occur, and implement solutions expeditiously. The system also allows the creation of calibration sets that can be used to prevent drift in the scores as signed to questions. This is especially useful when scoring large numbers of responses to a question (e.g., more than 30,000 responses per question in the state NAEP). In addition, the image-based scoring system allows all responses to a particular exercise to be scored continuously until the item is finished, thereby improving the validity and reliability of scorer judgments.

The reliability of scoring is monitored during the coding process through (1) backreading, where table leaders review about 10 percent of each scorer's work to confirm a consistent application of scoring criteria across a large number of responses and across time; (2) daily calibration exercises to reinforce the scoring criteria after breaks of more than 15 minutes; and (3) a second scoring of 25 percent of the items appearing only in the main national assessment and 6 percent of the items appearing in both the main national and state assessments (and a comparison of the two scores to give a measure of interscorer reliability). To monitor agreement across years, a random sample of 20–25 percent of responses from previous assessments (for identical items) is systematically interspersed among current responses for rescoring. If necessary, current assessment results are adjusted to account for any differences.

To test scoring reliability, constructed-response item score statistics are calculated for the portion of responses that are scored twice. Cohen's Kappa is the reliability estimate used for dichotomized items and the intraclass correlation coefficient is used as the index of reliability for nondichotomized items. Scores are also constructed for items that are rescored in a later assessment. For example, some 2007 reading and mathematics items were rescored in 2009.

Editing. The first phase of data editing takes place during the keying or scanning of the survey instruments. Machine edits verify that each sheet of each document is present and that each field has an appropriate value. The edit program checks each booklet number against the session code for appropriate session type, the school code against the control system record, and other data fields on the booklet cover for valid ranges of values. It then checks each block of the document for validity, proceeding through the items within the block. Each piece of input data is checked to verify that it is of an acceptable type, that the value falls within a specified range of values, and that it is consistent with other data values. At the end of this process, a paper edit listing of data errors is generated for nonimage and key-entered documents. Image-scanned items requiring correction are displayed at an online editing terminal.

In the second phase of data editing, experienced editing staff review the errors detected in the first phase, compare the processed data with the original source document, and indicate whether the error is correctable or noncorrectable per the editing specifications. Suspect items found to be correct as stated, but outside the edit specifications, are passed through modified edit programs. For nonimage and key-entered documents, corrections are made later via key-entry. For image-processed documents, suspect items are edited online. The edit criteria for each item in question appear on the screen along with the item, and corrections are made immediately. Two different people view the same suspect item and operate on it separately; a "verifier" ensures that the two responses are the same before the system accepts that item as correct.

For assessment items that must be paper-scored rather than scored using the image system (as was the case for some mathematics items in the 1996 NAEP), the score sheets are scanned on a paper-based scanning system and then edited against tables to ensure that all responses were scored with only one valid score and that only raters qualified to score an item were allowed to score it. Any discrepancies are flagged and resolved before the data from that scoring sheet are accepted into the scoring system.

In addition, a count-verification phase systematically compares booklet IDs with those listed in the NAEP administration schedule to ensure that all booklets expected to be processed were actually processed. Once all corrections are entered and verified, the corrected records are pulled into a mainframe data set and then re-edited with all other records. The editing process is repeated until all data are correct.

Estimation Methods

Once NAEP data are scored and compiled, the responses are weighted according to the sample design and population structure and then adjusted for nonresponse. This ensures that students' representation in NAEP matches their actual proportion of the school population in the grades assessed. The analyses of NAEP data for most subjects are conducted in two phases: scaling and estimation. During the scaling phase, item response theory (IRT) procedures are used to estimate the measurement characteristics of each assessment question. During the estimation phase, the results of the scaling are used to produce estimates of student achievement (proficiency) in the various subject areas. Marginal maximum likelihood (MML) methodology is then used to estimate characteristics of the proficiency distributions. Estimates of student achievement are included in the NAEP database; estimates of other variables are not included.

Weighting. The weighting for the national and state samples reflects the probability of selection for each student in the sample, adjusted for school and student nonresponse. The weight assigned to a student's responses is the inverse of the probability that the student would be selected for the sample. Prior to 2002, poststratification was used to ensure that the weighting was representative of certain subpopulations corresponding to figures from the U.S. Census and the Current Population Survey (CPS).

Student base weights. The base weight assigned to a student is the reciprocal of the probability that the student would be selected for a particular assessment. This probability is the product of the following two factors:

- the conditional probability that the school would be selected, given the strata; and
- the conditional probability, given the school, that the student would be selected within the school.

Nonresponse adjustments of base weights. The base weight for a selected student is adjusted by two nonresponse factors. The first factor adjusts for sessions that were not conducted. This factor is computed separately within classes formed by the first three digits of strata (formed by crossing the major stratum and the first socioeconomic characteristic used to define the final stratum). Occasionally, additional collapsing of classes is necessary to improve the stability of the adjustment factors, especially for the smaller assessment components. The second factor adjusts for students who failed to appear in the scheduled session or makeup session. This nonresponse adjustment is completed separately for each assessment. For assessed students in the trend samples, the adjustment is made separately for classes of students based on subuniverse and modal grade status. For assessed students in the main samples, the adjustment classes are based on subuniverse, modal grade status, and race class. In some cases, nonresponse classes are collapsed into one class to improve the stability of the adjustment factors.

NIES Part II weighting. In the NIES Part II,. the school probability of selection is a function of three factors: NAEP selection, the probability of being retained for NIES Part II, and the number of AI/AN students in the NAEP sample per school. Nonresponse adjustments at the school level attempt to mitigate the impact of differential response by school type (public, private, and BIE), region, and estimated percentage enrollment of AI/AN students. For student weights, nonresponse adjustments take into account differential response rates based on student age (above age for grade level or not) and English language learner status. In order to partially counteract the negative impact of low private school participation, a poststratification adjustment is applied to the NIES Part II weights. The relative weighted proportions of students from public, private, and BIE schools, respectively, are adjusted to match those from the NIES Part I data. This not only ensured greater consistency between the findings of the two NIES components, but since the proportions of students are more reliably estimated from the NIES Part I data (which involved a far larger school sample

than Part II), this weight adjustment increases the accuracy and reliability of the NIES Part II results.

Scaling. For purposes of summarizing item responses, ETS developed a scaling technique that has its roots in IRT procedures and the theories of imputation of missing data.

The first step in scaling is to determine the percentage of students who give various responses to each cognitive, or subject-matter, question and each background question. For cognitive questions, a distinction is made between missing responses at the end of a block (i.e., missing responses after the last question the student answered) and missing responses before the last observed response. Missing responses before the last observed response are considered intentional omissions. Missing responses at the end of a block are generally considered "not reached" and treated as if the questions had not been presented to the student. In calculating response percentages for each question, only students classified as having been presented that question are used in the analysis. Each cognitive question is also examined for differential item functioning (DIF). DIF analyses identify questions on which the scores of different subgroups of students at the same ability level differ significantly.

Development of scales. Separate subscales are derived for each subject area. For the main assessments, the frameworks for the different subject areas dictate the number of subscales required. In the 2009 NAEP, five subscales were created for the main assessment in mathematics in grades 4 and 8 (one for each mathematics content strand), and three subscales were created for science (one for each field of science: Earth, physical, and life). A composite scale is also created as an overall measure of students' performance in the subject area being assessed (e.g., mathematics). The composite scale is a weighted average of the separate subscales for the defined subfields or content strands. For the long-term trend assessments, a separate scale is used for summarizing proficiencies at each age in mathematics and reading.

Within-grade vs. cross-grade scaling. The reading and mathematics main NAEP assessments were developed with a cross-grade framework, where the trait being measured was conceptualized as cumulative across the grades of the assessment. Accordingly, a single 0–500 scale was established for all three grades in each assessment. In 1993, however, NAGB determined that future NAEP assessments should be developed using within-grade frameworks and be scaled accordingly. This both removed the constraint that the trait being measured is cumulative and eliminated the need for overlap of questions across grades. Any questions that happen to be the same across grades are scaled separately for each grade, thus making it possible for common questions to function differently in the separate grades.

The 1994 history and geography assessments were developed and scaled within grade, according to NAGB's new policy. The scales were aligned so that grade 8 had a higher mean than grade 4 and grade 12 had a higher mean than grade 8. The 1994 reading assessment, however, retained a cross-grade framework and scaling. All three main assessments in 1994 used scales ranging from 0 to 500.

The 2008 long-term trend assessments remained cross-age, using a 0–500 scale. The 2009 main science assessment was developed within-grade, but adopted new scales ranging from 0 to 300. The 2005 main assessment in mathematics continued to use a cross-grade framework with a 0–500 scale in grades 4 and 8, but used a 0-300 within-grade scale. In 1998, reading, writing and civics assessments were scaled within-grade.

Linking of scales. Before 2002, results for the main state assessments were linked to the scales for the main national assessments, enabling state and national trends to be studied. Equating the results of the state and national assessments depended on those parts of the main national and state samples that represented a common population: (1) the state comparison sample—students tested in the national assessment who come from the jurisdictions participating in the state NAEP; and (2) the state aggregate sample—the aggregate of all students tested in the state NAEP. Since 2002, the national sample has been a superset of the state samples (except in those states that do not participate). Thus, equating is not necessary.

Imputation. **U**ntil the 2002 NAEP assessment, no statistical imputations were generated for missing values in the teacher, school, or SD/ELL questionnaires, or for missing answers to cognitive questions. Most answers to cognitive questions are missing by design. For example, 8th-grade students being assessed in reading are presented with, on average, 21 of the 110 assessment items. Whether any given student gets any of the remaining 89 individual questions right or wrong is not something that NAEP imputes. However, since 1984, multiple imputation

techniques have been used to create plausible values. Once created, subsequent users can analyze these plausible values with common software packages to obtain NAEP results that properly account for NAEP's complex item sampling designs.

Because no student takes even a quarter of the questions in an assessment, individual scores cannot be calculated. Trying to use partial scores based on the small proportion of the assessment to which any given student is exposed would lead to biased results for groups scores due to an inherently large component of measurement error. NAEP developed a process of group score calculation in order to get around the unreliability and noncomparability of NAEP's partial test forms for individuals. NAEP estimates group score distributions using MML estimation, a method that calculates group score distributions based directly on each student's responses to cognitive questions, not on summary scores for each student. As a result, the unreliability of individual-level scores does not decrease NAEP's accuracy in reporting group scores. The MML method does not employ imputations of answers to any questions or of scores for individuals.

Imputation is performed in three stages. The first stage requires estimating IRT parameters for each cognitive question. The second stage results in MML estimation of a set of regression coefficients that capture the relationship between group score distributions and nearly all the information from the variables in the teacher, school, or SD/ELL questionnaires, as well as geographical, sample frame, and school record information. The third stage involves the imputation that is designed to reproduce the group-level results that could be obtained during the second stage.

NAEP's imputations follow Rubin's (1987) proposal that the imputation process be carried out several times, so that the variability associated with group score distributions can be accurately represented. NAEP estimates five plausible values for each student. The five plausible values are calculated using the regression coefficients estimated in the second stage. Each plausible value is a random selection from the joint distribution of potential scale scores that fit the observed set of response for each student and the scores for each of the groups to which each student belongs. Estimates based on plausible values are more accurate than if a single (necessarily partial) score were to be estimated for each student and averaged to obtain estimates of subgroup performances. Using the plausible values eliminates the need for secondary analysts to have access to specialized MML software and ensures that the estimates of average performance of groups and estimates of variability in those averages are accurate.

Recent Changes

Several important changes have been implemented since 1990.

- Beginning with the 1990 mathematics assessment, NAGB established three reporting levels for reporting NAEP results: basic, proficient, and advanced.

- In 1990, state assessments were added to NAEP. The 1990 to 1994 assessments are referred to as trial state assessments.

- In 1992, a generalized partial-credit model (GPCM) was introduced to develop scales for the more complex constructed-response questions. The GPCM model permits the scaling of questions scored according to multipoint rating schemes.

- In 1993, NAGB determined that future NAEP assessments should have within-grade frameworks and scales. The 1994 main history and geography assessments followed this new policy, as did the 1996 main science assessment, and the 1998 writing assessment. Mathematics and reading in the main NAEP will continue to have cross-grade scales until further action by NAGB (and a parallel change in the trend assessment), except for mathematics at grade 12, which was removed from cross-grade scales and reported in a within-grade scale in 2005.

- In 1994, the new image-based scoring system virtually eliminated paper handling during the scoring process. This system also permits scoring reliability to be monitored online and recalibration methods to be introduced.

- The 1996 main NAEP included new samples for the purpose of studying greater inclusion of SD/LEP students and obtaining data on students eligible for advanced mathematics or science sessions.

- In 1997, there was a probe of student performance in the arts.

- New assessment techniques included: open-ended items in the 1990 mathematics assessment; primary trait, holistic, and writing mechanics scoring procedures in the 1992 writing assessment; the use of calculators in the 1990, 1992, 1996, and 2000 mathematics assessments;

a special study on group problem solving in the 1994 history assessment; and a special study in theme blocks in the 1996 mathematics and science assessments.

- Beginning in 1998, testing accommodations were provided in the NAEP reading assessments; in this transition to a more inclusive NAEP, administration procedures were introduced that allowed the use of accommodations (e.g., extra time, individual rather than group administration) for students who required them to participate. During this transition period, reading results in 1998 were reported for two separate samples: one in which accommodations were not permitted and one in which accommodations were permitted. Beginning in 2002, accommodations were permitted for all reading administrations.

- In 1999, NAGB discontinued the long-term trend assessment in writing for technical reasons. More recently, NAGB decided that changes were needed to the design of the science assessment and, given recent advances in the field of science, to its content. As a result, the science long-term trend assessment was not administered in 2003-04.

- With the expansion and redesign of NAEP under the No Child Left Behind Act, NAEP's biennial state-level assessments are being administered by contractor staff (not local teachers). The newly redesigned NAEP has four important features. First, NAEP is administering tests for different subjects (such as mathematics, science, and reading) in the same classroom, thereby simplifying and speeding up sampling, administration, and weighting. Second, NAEP is conducting pilot tests of candidate items for the next assessment and field tests of items for precalibration in advance of data collection, thereby speeding up the scaling process. Third, NAEP is conducting bridge studies, administering tests both under the new and the old conditions, thereby providing the possibility of linking old and new findings. Finally, NAEP is adding additional test questions at the upper and lower ends of the difficulty spectrum, thereby increasing NAEP's power to measure performance gaps.

- Beginning in 2002, the NAEP national sample for main national assessment was obtained by aggregating the samples from each state, rather than by obtaining an independently selected national sample. Prior to 2002, separate samples were drawn for the NAEP main national and state assessments.

- In 2002, TUDA began assessing performance in five large urban districts with reading and writing assessments. TUDA continued in 2003 in nine large urban districts with reading and mathematics and in 2005 in 10 large urban districts with reading, mathematics, and science.

- Beginning with the 2003 NAEP, each state must have participation from at least 85 percent—instead of 70 percent—of the schools in the original sample in order to have its results reported.

- In 2003 and 2005, Puerto Rico participated in the NAEP assessment of mathematics. However, Puerto Rico was excused from the NAEP assessment of reading in English because Spanish is the language of instruction in Puerto Rico. NCES also administered the 2007 mathematics assessment in Puerto Rico.

- In 2004, several changes were implemented to the NAEP long-term trend assessments to reflect changes in NAEP policy, maintain the integrity of the assessments, and increase the validity of the results obtained. The changes to the assessment instruments include: removal of science items; inclusion of students with disabilities and English language learners; replacement of items that used outdated contexts; creation of a separate background questionnaire; elimination of "I don't know" as a response option for multiple-choice items; and use of assessment booklets that pertain to a single subject area (whereas in the past, a single assessment booklet may have contained both reading and mathematics items).

- In 2005, NAGB introduced changes in the NAEP mathematics framework for grade 12 in both the assessment content and administration procedures. One of the major differences between the 2005 assessment and previous assessments at grade 12 is the five content areas were collapsed into four areas, with geometry and measurement being combined. In addition, the assessment included more questions on algebra, data analysis, and probability to reflect changes in high school mathematics standards and coursework. The overall average mathematics score in 2005 was set at 150 on a 0–300 scale.

- In 2006, economics was assessed at grade 12 for the first time. A within-grade scale was developed, with the overall average economics score in 2006 set at 150 on a 0–300 scale.

- In 2009, TUDA was expanded to 18 large urban districts, assessing reading, mathematics and science. In addition, 11 states were assessed in reading and mathematics at grade 12 on a trial basis.

- In 2009, interactive computer tasks in science were administered online at grades 4, 8, and 12. These tasks consisted of simulations for the students to draw inferences and conclusions about a problem.

Future Plans

The next trend assessment will be administered in 2012, and then every 4 years thereafter. Main assessments are scheduled for annual administration. Reading and mathematics are assessed every 2 years in odd-numbered years; science and writing are scheduled to be assessed every 4 years (in the same years as reading and mathematics, but alternating with each other); and other subjects are assessed at the national level in even-numbered years. Writing will be assessed online in 2011 to a national sample of 4^{th} and 8^{th} graders.

5. DATA QUALITY AND COMPARABILITY

As the Nation's Report Card, NAEP must report accurate results for populations of students and subgroups of these populations (e.g., Black, Hispanic, or other race/ethnicity, or students attending nonpublic schools). Although only a very small percentage of the student population in each grade is assessed, NAEP estimates are accurate because they depend on the absolute number of students participating, not on the relative proportion of students.

Every activity in NAEP assessments is conducted with rigorous quality control, contributing both to the quality and comparability of the assessments and their results. All questions undergo extensive reviews by subject-area and measurement specialists, as well as careful scrutiny to eliminate any potential bias or lack of sensitivity to particular groups. The complex process by which NAEP data are collected and processed is monitored closely. Although each participating state is responsible for its own data collection for the main state NAEP, Westat ensures uniformity of procedures across states through training, supervision, and quality control monitoring.

With any survey, however, there is the possibility of error. The most likely sources of error in NAEP are described below.

Sampling Error

Two components of uncertainty in NAEP assessments are accounted for in the variability of statistics based on scale scores: (1) the uncertainty due to sampling only a small number of students relative to the whole population; and (2) the uncertainty due to sampling only a relatively small number of questions. The variability of estimates of percentages of students having certain back-ground characteristics or answering a certain cognitive question correctly is accounted for by the first component alone.

Because NAEP uses complex sampling procedures, a jackknife replication procedure is used to estimate standard errors. While the jackknife standard error provides a reasonable measure of uncertainty about student data that can be observed without error, each student in NAEP assessments typically responds to so few questions within any content area that the scale score for the student would be imprecise. It is possible to describe the performance of groups and subgroups of students because, as a group, all students are administered a wide range of items.

NAEP uses MML procedures to estimate group distributions of scores. However, the underlying imprecision that makes this step necessary adds an additional component of variability to statistics based on NAEP scale scores. This imprecision is measured by the imputed variance, which is estimated by the variance among the plausible values drawn from each student's posterior distribution of possible scores. The final estimate of the variance is the sum of the sampling variance and the measurement variance.

Nonsampling Error

While there is the possibility of some coverage error in NAEP, the two most likely types of nonsampling error are nonresponse error due to nonparticipation and measurement error due to instrumentation defects (described below). The overall extent of nonsampling error is largely unknown.

Coverage error. In NAEP, coverage error can result either from the sampling frame of schools being incomplete or from the schools' failure to include all the students on the lists from which grade or age samples are drawn. For the 2009 NAEP, the 2008

school list maintained by CCD supplied the names of the regular public schools, BIE schools, and DoDEA schools. This list, however, did not include schools that opened between 2008 and the time of the 2009 NAEP. To be sure that students in new public schools were represented, each sample district in NAEP was asked to update lists of schools with newly eligible schools.

Catholic and other nonpublic schools in the 2009 NAEP were obtained from the PSS. PSS uses a dual-frame approach. The list frame (containing most private schools in the country) is supplemented by an area frame (containing additional schools identified during a search of randomly selected geographic areas around the country). Coverage of private schools in the PSS is very high. (See chapter 3.)

Nonresponse error.
Unit nonresponse. In the 2009 reading and mathematics assessments, all 52 states and jurisdictions[3] met participation rate standards at both grade 4 and grade 8. The national school participation rates for public and private schools combined were 97 percent at grades and grade 8. Student participation rates were 95 percent at grade 4 and 93 percent at grade 8. Participation rates needed to be 70 percent or higher to report results separately for private schools. While the participation rate for private schools did meet the standard in 2009, it did not always meet the standard in previous assessment years. See table 11 for more details.

In the 2007 reading and mathematics assessments, all 52 states and jurisdictions[4] met participation rate standards at both grades 4 and 8. The national school participation rates for public and private schools combined were 98 percent at grade 4 and 97 percent at grade 8. Student participation rates were 95 percent at grade 4 and 92 percent at grade 8. Participation rates needed to be 70 percent or higher to report results separately for private schools.

In the 2005 reading and mathematics assessments at grade 12, participation standards were met for public schools but not for private schools. At the student level, response rates at grade 12 fell below 85 percent for students in both public and private schools. A nonresponse bias analysis showed significant differences between responding and nonresponding public school students in terms of gender, race/ethnicity, age, and English language learner identification. Although the differences are quite small, it is unlikely that nonresponse weighting adjustments completely accounted for these differences.

In the 2008 trend assessments, private school participation rate at age 17 was 61 percent, below the standard for reporting. However, Catholic school participation rates at all three ages (88, 94, and 76 percent at ages 9, 13, and 17, respectively) met the reporting standards.

In the 2007 NIES Part II, questionnaires were completed by about 10,400 grade 4 students from 1,700 schools and 11,300 grade 8 students from 1,800 schools. Also responding to the survey were about 3,000 grade 4 teachers, 4,600 grade 8 teachers, 1,700 grade 4 school administrators and 1,800 grade 8 school administrators associated with these students. Some school administrators responded for both grades 4 and 8. The weighted student response rates were 85 percent at grade 4 and 82 percent at grade 8. The weighted school response rates were 88 percent at grade 4 and 90 percent at grade 8.

In the 2005 NIES Part II, questionnaires were completed by about 2,600 grade 4 students and 2,500 grade 8 students at approximately 480 schools. Also responding to the survey were about 480 grade 4 teachers, 820 grade 8 teachers, 240 grade 4 principals, and 230 grade 8 principals associated with these students. Some principals responded for both grades 4 and 8. The weighted student response rates were 95 percent at grade 4 and 91 percent at grade 8. The weighted school response rates were 87 percent at grade 4 and 93 percent at grade 8.

In the 2004 long-term trend reading and mathematics assessments, the overall response rate (the product of the weighted school participation rate before substitution and the weighted student participation rate) fell below the NCES reporting target of 85 percent for ages 13 and 17 at the school level and for age 17 at the student level. At age 13, a bias was found for private schools, as a greater proportion of nonresponses were from other private schools than from Catholic schools. In addition, nonrespondent schools in the long-term trend assessment had a lower percentage of Black students than participating schools. Likewise, at age 17, private schools were disproportionately less likely to participate, and within private schools, Catholics and Conservative Christian schools had higher participation rates than other private schools. Nonrespondent schools

[3] It includes 50 states, District of Columbia, and DoDEA.
[4] It includes 50 states, District of Columbia, and DoDEA.

Table 11. Weighted school, student, and overall response rates for selected NAEP national assessments, by assessment and grade: 2006-2009

Assessment and grade	School participation[1]		Student participation	
	Student weighted	School weighted	Student weighted	Overall participation
2009 Mathematics				
Grade 4	97	91	95	92
Grade 8	97	87	93	90
2009 Reading				
Grade 4	97	91	95	92
Grade 8	97	87	93	90
2009 Science				
Grade 4	97	91	95	92
Grade 8	97	87	93	90
Grade 12	83	79	80	66
2008 Trend				
Age 9	96	91	95	91
Age 13	95	89	94	89
Age 17	90	85	88	79
2007 Writing				
Grade 8	97	87	92	90
Grade 12	89	83	80	71
2007 Reading				
Grade 4	98	92	95	93
Grade 8	97	87	92	90
2007 Mathematics				
Grade 4	98	92	95	93
Grade 8	97	87	92	90
2006 Economics				
Grade 12	79	78	73	58
2006 Civics				
Grade 4	92	86	95	88
Grade 8	93	86	92	85
Grade 12	79	78	72	57
2006 U.S. history				
Grade 4	91	88	95	87
Grade 8	91	85	92	84
Grade 12	80	80	73	59

[1] Participation rates do not include substitutions.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2009 Mathematics, Reading and Science Assessments, 2008 Trend Assessment, 2007 Writing, Reading and Mathematics Assessments, 2006 Economics, Civics, and U.S. history Assessments.

also had a slightly higher percentage of Asian students than participating schools at age 17. At the student level at age 17, some bias was shown for race/ethnicity, free lunch eligibility, and disability status.

Item nonresponse. Specific information about nonresponse for particular items is available in NAEP summary data tables on the Web http://nces.ed.gov/nationsreportcard/naepdata/.

Measurement error. Nonsampling error can result from the failure of the test instruments to measure what is being taught and, in turn, what is being learned by students. For example, the instruments may contain ambiguous definitions and/or questions that lead to different interpretations by students. Additional sources of measurement error are the inability or unwillingness of students to give correct information and errors in the recording, coding, or scoring of data.

To assess the quality of the data in the final NAEP database, survey instruments are selected at random and compared, character by character, with their records in the final database. As in past years, the 2000 NAEP data-base was found to be more than accurate enough to support analyses.

The observed error rates for the 2000 NAEP were comparable to those of past assessments. Error rates ranged from 8 errors per 10,000 responses for the Teacher Survey questionnaire to 44 errors per 10,000 responses for the School Characteristics and Policies Survey questionnaire.

Revised results. Following the 1994 NAEP assessment, two technical problems were discovered in the procedures used to develop the scale and achievement levels for the 1990 and 1992 mathematics assessments. These errors affected the mathematics scale scores reported for 1992 and the achievement-level results reported for 1990 and 1992.

NCES and NAGB evaluated the impact of these errors and subsequently reanalyzed data and reported the revised results from both mathematics assessments. The revised results for 1990 and 1992 are presented in the 1996 mathematics reports. For more detail on these problems, see *The NAEP 1996 Technical Report* (Allen, Carlson, and Zelenak 1999) and the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics* (Allen et al. 1997).

There were also problems related to reading scale scores and achievement levels. These errors affected the 1992 and 1994 NAEP reading assessment results. The 1992 and 1994 reading data have been reanalyzed and reissued in revised reports. For more information, refer to *The NAEP 1994 Technical Report* (Allen, Kline, and Zelenak 1996) and the *Technical Report of the NAEP 1994 Trial State Assessment in Reading* (Mazzeo, Allen, and Kline 1995).

Data Comparability

NAEP allows reliable comparisons between state and national data for any given assessment year. By linking scales across assessments, it is possible to examine short-term trends for data from the main national and state NAEP and long-term trends for data from the long-term trend NAEP.

Main national vs. main state comparisons. NAEP data are collected using a closely monitored and standardized process, which helps ensure the comparability of the results generated from the main national and state assessments. The main national NAEP and main state NAEP use the same assessment booklets, and, since 2002, they have been administered in the same sessions using identical procedures.

Short-term trends. Although the test instruments for the main national assessments are designed to be flexible and thus adaptable to changes in curricular and educational approaches, they are kept stable for shorter periods (up to 12 years or more) to allow analysis of short-term trends. For example, through common questions, the 1996 main national assessment in mathematics was linked to both the 1992 and 1994 assessments.

For 2005, NAGB adopted a new mathematics framework for grade 12 to reflect changes in high school standards and coursework. In addition, changes were made in booklet design and calculator-use policy for the one-third of the assessment in which calculators were allowed. As a result of these changes, the 2005 results could not be placed on the previous NAEP scale and are not compared to results from previous years.

Long-term trends. In order to make long-term comparisons, the long-term trend NAEP uses different samples than the main national NAEP. Unlike the test instruments for the main NAEP, the long-term instruments in mathematics and reading have remained relatively unchanged from those used in previous assessments. The 2004 trend instruments were almost identical to those used in the 1970s. The trend NAEP allows the measurement of educational

progress since 1971 in reading and 1973 in math ematics. For more detail on the linking of scales in the trend NAEP, see "Scaling" in section 4 above.

The long-term trend assessment was updated in several ways in 2004 (e.g., inclusion of SD/ELL students). To ensure the comparability of the new assessment and the previous assessments, a bridge study was performed.

Linking to non-NAEP assessments. Linking results from the main state assessments to those from the main national assessments has encouraged efforts to link NAEP assessments with non-NAEP assessments.

Linking to state assessment. NAEP data can be used to map state proficiency standards in reading and mathematics onto the appropriate NAEP scale. The mapping exercise was carried out for data from the 2004–05 and 2006-07 academic years at both grades 4 and 8. For each of the four subject and grade combinations, the NAEP score equivalents to the states' proficiency standards vary widely, spanning a range of 60 to 80 NAEP score points. Although there is an essential ambiguity in any attempt to place state standards on a common scale, the ranking of the NAEP score equivalents to the states' proficiency standards offers an indicator of the relative stringency of those standards. There are plans to do this mapping for the 2008-09 school year also.

There is a strong negative correlation between the proportions of students meeting the states' proficiency standards and the NAEP score equivalents to those standards, suggesting that the observed heterogeneity in states' reported percents proficient can be largely attributed to differences in the stringency of their standards. There is, at best, a weak relationship between the NAEP score equivalents to the states' proficiency standards and the states' average scores on NAEP. Finally, most of the NAEP score equivalents fall below the cut-point corresponding to the NAEP proficient level, and many fall below the cut-point corresponding to the NAEP basic level.

These results should be employed cautiously, as differences among states in apparent stringency can be due, in part, to reasonable differences in the assessment frameworks, the types of item formats employed, and the psychometric characteristics of the tests. Moreover, there is some variation among states in the proportion of NAEP sample schools that could be employed in the analysis.

Linking to the International Assessment of Educational Progress (IAEP). In 1992, results from the 1992 NAEP assessment in mathematics in grade 8 were successfully linked to those from IAEP of 1991. Sample data were collected from U.S. students who had been administered both instruments. The relation between mathematics proficiency in the two assessments was modeled using regression analysis. This model was then used as the basis for projecting IAEP scores from non-U.S. countries onto the NAEP scale. *The relation between the IAEP and NAEP assessments was relatively strong and could be modeled well. The results, however, should be considered only in the context of the similar construction and scoring of the two assessments. Further studies should be initiated cautiously, even though the path to linking assessments is now better understood.*

Linking to TIMSS. The success in linking NAEP to the IAEP sparked an interest in linking the results from the 1996 NAEP assessments in mathematics and science in grade 8 to those from the Third International Mathematics and Science Study (TIMSS) of 1995. The data from this study became available at approximately the same time as the1996 NAEP data for mathematics and science. Because the two assessments were conducted in different years and no students responded to both assessments, the regression procedure that linked NAEP and IAEP assessments could not be used. The results from grade 8 NAEP and TIMSS assessments were instead linked by matching their score distributions. A comparison of the linked results with actual results from states that participated in both assessments suggested that the link was working acceptably. *The results from U.S. students were linked to those of their academic peers in more than 40 other countries. As with the IAEP linked results, these results should be used cautiously.*

A second study attempted to link the 2000 grade 8 NAEP assessments in mathematics and science to the 1999 grade 8 TIMSS (which also assessed mathematics and science). The primary linkage used a projection method, which drew data from a sample of students to whom both assessments were administered. The linkage found that the projections were substantially off the mark. A secondary linkage, based on nationally reported numbers using a statistical moderation approach, provided a fairly weak linkage; the moderation linkage did a decent job of projecting TIMSS scores from NAEP scores in the 12 states that participated in both studies, but failed to predict the TIMSS score in the linking sample.

The analyses showed that the TIMSS assessments functioned differently in the linking sample than they did in the national and state samples. A recent study (Phillip 2009) shows that it is possible to make comparisons between TIMSS 2007 and NAEP 2007. For more details, please refer to *The Second Derivative: International Benchmarks in Mathematics for U.S. States and School Districts* (Phillip 2009).

Comparisons with TIMSS. Studies were undertaken to compare the content of two fourth- and eighth-grade assessments in mathematics and science: the NAEP 2000 assessment and the TIMSS 2003 assessment. The comparison study drew upon information provided by the developers of the assessments, as well as data obtained from an expert panel convened to compare the frameworks and items from the two assessments on various dimensions.

For science, the content comparisons between NAEP and TIMSS reveal some key differences in the topics covered, grade-level correspondence, and the characteristics of the item pools on other dimensions. All of these factors together may result in differences in student performance, and it is important to consider these differences when interpreting the results from the different assessments.

Differences in the science content included in each assessment can be seen at both the framework level and in the pool of items developed based on these frameworks. Even in content areas where there is considerable overlap of the frameworks (such as life science and Earth science), a closer examination of the topics and specific objectives covered by the items in each assessment reveals some important differences. In comparison to NAEP, whose framework was developed in the context of the U.S. system, the TIMSS framework reflects a consensus across many countries. Some of the differences in curricula across these countries are reflected in the frameworks and in the differences in content of the two assessments. In particular, the inclusion in TIMSS of separate content areas in chemistry, physics, and environmental science results in broader topic coverage in some areas. While there is a considerable overlap in the topics included in some content areas, the items included in each assessment place different emphases at the topic level. In addition, the "hands-on" tasks in NAEP provide complementary information to the pencil-and-paper portions of both assessments, enabling the measurement of student performance in this area of knowing and doing science.

With respect to mathematics, a comparison of the frameworks revealed considerable agreement on the general boundaries and basic organization of mathematics content, with both assessments including five main content areas corresponding to traditional mathematics curricular areas: number, measurement, geometry, data, and algebra. Both the NAEP and TIMSS frameworks also include dimensions that define a range of cognitive skills and processes that overlap the two assessments. Despite these apparent similarities at the broadest level, a closer examination of the items in each assessment reveals different emphases at the topic and subtopic levels, as well as some differences in grade-level expectations across mathematics topics.

Comparisons with PIRLS. In 2003, NCES released results for both the 2001 Progress in International Reading Literacy Study (PIRLS) fourth-grade assessment and the 2002 NAEP fourth-grade reading assessment. In anticipation of questions about how these two assessments compare, NCES convened an expert panel to compare the content of the PIRLS and NAEP assessments and determine if they are measuring the same construct. This involved a close examination of how PIRLS and NAEP define reading, the texts used as the basis for the assessments, and the reading processes required of students in each. The comparison suggests that there is a great deal of overlap in what the two assessments are measuring. While they do seem to define and measure the same kind of reading, PIRLS is an easier assessment than NAEP, with more text-based tasks and shorter, less complex reading passages. The similarities and differences between the two are discussed below.

The comparison revealed that, overall, the NAEP and PIRLS reading assessments are quite similar. Both define reading similarly, as a constructive process. Both use high-quality reading passages and address similar purposes for which young children read (for literary experience and information). Both call for students to develop interpretations, make connections across text, and evaluate aspects of what they have read. Finally, both have a similar distribution of multiple-choice and constructed-response items: in each, about half of the items are constructed-response items.

While the two assessments have similar definitions of reading and assess many of the same aspects of it, a closer look at how the domain is operationalized by each revealed some important differences. NAEP places more emphasis than PIRLS on having students taking what they have read and connecting it to other readings or knowledge. PIRLS places a greater

emphasis than NAEP on text-based reading skills and interactions, including items that ask students to locate information in the text, make text-based inferences and interpretations, and evaluate aspects of the text.

The PIRLS reading passages are, on average, about half the length of the NAEP reading passages. PIRLS readability formulas indicate that the passages used in PIRLS are less complex than those used in NAEP. The classification of items also revealed differences in how the two frameworks function. The panel had an easier time classifying PIRLS and NAEP items by the PIRLS framework categories than by the NAEP framework categories. For more information on the similarities and differences between PIRLS and NAEP, see *A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments* (Binkley and Kelly 2003).

Comparisons with the International Association for the Evaluation of Educational Achievement's (IEA) Reading Literacy Study. The picture of American students' reading proficiency provided by NAEP assessments is less optimistic than that indicated by the IEA Reading Literacy Study. This can be explained by the following:

(1) *The basis for reporting differs considerably between the two assessments.* With the IEA study, students are compared against other students and not against a standard set of criteria on knowledge, as in NAEP. Much of NAEP reporting is based on comparisons between actual student performance and desired performance (what students are expected to do).

(2) *NAEP and IEA assess different aspects of reading.* More than 90 percent of the IEA items assess tasks covered in only 17 percent of NAEP items. Furthermore, virtually all of the IEA items are aimed solely at literal comprehension and interpretation, while such items make up only one-third of NAEP reading assessments.

(3) *NAEP and IEA differ in what students must do to demonstrate their comprehension.* More interpretive and higher level thinking is required to reach the advanced level in NAEP than in the IEA study. Also, NAEP requires students to generate answers in their own words much more frequently than does the IEA study. Moreover, the IEA test items do not cover the entire expected ability range. Many American students answer every IEA item correctly, making it impossible to distinguish between the abilities of students in the upper range. In contrast, the range of item difficulty on NAEP reading assessments exceeds the ability of most American students, so differences in the abilities of students in the upper range can be distinguished easily.

Despite the differences between these two assessments, there is a high probability that, if students from other countries were to take NAEP, the rank ordering or relative performance of countries would be about the same as in the IEA findings. This assumption is based on the theoretic underpinnings of item response theory and its application to the test scaling used for both the IEA Reading Literacy Study and the NAEP reading assessment.

6. CONTACT INFORMATION

For content information on NAEP, contact:

Peggy Carr
Phone: (202) 502-7321
E-mail: peggy.carr@ed.gov

Mailing Address:
National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

7. METHODOLOGY AND EVALUATION REPORTS

General

Allen, N.L., Carlson, J.E., and Zelenak, C.A. (1999). *The NAEP 1996 Technical Report* (NCES 1999-452). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Allen, N.L., Donoghue, J.R., and Schoeps, T.L. (2001). *The NAEP 1998 Technical Report* (NCES 2001-509). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Allen, N.L., Jenkins, F., Kulick, E., and Zelenak, C. A. (1997). *Technical Report of the NAEP 1996 State Assessment Program in Mathematics* (NCES 97-

951). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Allen, N.L., Kline, D.L., and Zelenak, C.A. (1996). *The NAEP 1994 Technical Report* (NCES 97-897). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Allen, N.L., Swinton, S.S., Isham, S.P., and Zelenak, C.A. (1998). *Technical Report: NAEP 1996 State Assessment Program in Science* (NCES 98-480). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Alt, M.N., and Bradby, D. (1999). *Procedures Guide for Transcript Studies* (NCES Working Paper 1999-05). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Braun, H., Jenkins, F., and Grigg, W. (2006). *Comparing Private Schools and Public Schools Using Hierarchical Linear Modeling* (NCES 2006-461). U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office.

Calderone, J., King, L.M., and Horkay, N. (1997). *The NAEP Guide: A Description of the Content and Methods of the 1997 and 1998 Assessments* (NCES 97-990). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Johnson, C. (2004). *Nation's Report Card: An Overview of NAEP* (NCES 2004-552). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Jonhson, E.G., and Carlson, J.E. (1994). *NAEP 1992 Technical Report* (NCES 94-490). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Jonhson, E.G., Mazzeo, J., and Kline, D.L. (1994). *Technical Report of the NAEP 1992 Trial State Assessment in Reading* (NCES 94-472). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Lutkus, A.D., Mazzeo, J., Zhang, J., Jerry, L., Barton, K.E., and Zenisky, A. (2003). *Including Special-Needs Students in the NAEP 1998 Reading Assessment, Part I, Comparison of Overall Results With and Without Accommodations* (NCES 2003-467). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Mazzeo, J., Allen, N.L., and Kline, D.L. (1995). *Technical Report of the NAEP 1994 Trial State Assessment in Reading* (NCES 96-116). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Moran, R., and Rampey, B. (2008). *National Indian Education Study 2007 - Part II: The Educational Experiences of American Indian and Alaska Native Students in Grades 4 and 8* (NCES 2008-458). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Patricia, D. (2005). *The Nation's Report Card: An Introduction to The National Assessment of Educational Progress (NAEP)* (NCES 2005-454). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Perie, M., Vanneman, A., and Goldstein, A. (2005). *Student Achievement in Private Schools: Results From NAEP 2000–2005* (NCES 2006-459). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, D.C.

Persky, H., Brent, S.A., and Janice, A. (2003). *Assessing the Arts: Selected NAEP Tasks and Scoring Guides for Grades 4 and 12 1997 Field Test. Dance, Music, Theatre, and Visual Art* (NCES 2003-452). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Phillips, G., and Goldstein, A. (1996). *Technical Issues in Large-Scale Performance Assessment* (NCES 96-802). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Roey, S., Caldwell, N., Rust, K., Blumstein, E., Krenzke, T., and Legum, S. (2001). *The 1998 High School Transcript Study User's Guide and Technical Report* (NCES 2001-477). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Roey, S., Caldwell, N., Rust, K., Hicks, L., Lee, J., Perkins, R., Blumstein, E., and Brown, J. (2005). *The High School Transcript Study: The 2000 High School Transcript Study User's Guide and Technical*

Report (NCES 2005-483). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2007). *Mapping 2005 State Proficiency Standards Onto the NAEP Scales* (NCES 2007-482). Washington, DC.

White, S. (1994). *Overview of NAEP Assessment Frameworks* (NCES 94-412). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Uses of Data

Phillips, G. (1993). *Interpreting NAEP Scales* (NCES 93-421). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Vanneman, A. (1997). *Focus on NAEP: New Software Makes NAEP Data User Friendly* (NCES 97-045). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Vanneman, A., Hamilton, L., Baldwin Anderson, J., and Rahman, T. (2009). *Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress*, (NCES 2009-455). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Survey Design

Bay, L., Chen, L., Hanson, B.A., Happel, J., Kolen, M.J., Miller, T., Pommerich, M., Sconing, J., Wang, T., and Welch, C. (1997). *ACT's NAEP Redesign Project: Assessment Design Is the Key to Useful and Stable Assessment Results* (NCES Working Paper 97-39). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Bock, D.R., and Zimowski, M.F. (2003). *NAEP Validity Studies: Feasibility Studies of Two-Stage Testing in Large-Scale Educational Assessment: Implications for NAEP* (NCES Working Paper 2003-14). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Chromy, J.R. (2003). *NAEP Validity Studies: The Effects of Finite Sampling on State Assessment Sample Requirements* (NCES Working Paper 2003-17). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Durán, R.P. (2003). *NAEP Validity Studies: Implications of Electronic Technology for the NAEP Assessment* (NCES Working Paper 2003-16). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Hedges, L.V., Konstantopoulos, S., and Thoreson, A. (2003). *NAEP Validity Studies: Computer Use and Its Relation to Academic Achievement in Mathematics, Reading, and Writing* (NCES Working Paper 2003-15). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Hedges, L.V., and Vevea, J.L. (2003). *AEP Validity Studies: A Study of Equating in NAEP* (NCES Working Paper 2003-13). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Jaeger, R.M. (2003). *NAEP Validity Studies: Reporting the Results of the National Assessment of Educational Progress* (NCES Working Paper 2003-11). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Jakwerth, P.R., Stancavage, R.B., and Reed, E.D. (2003). *NAEP Validity Studies: An Investigation of Why Students Do Not Respond to Questions* (NCES Working Paper 2003-12). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Johnson, E.G., Lazer, S., and O'Sullivan, C.Y. (1997). *NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress* (NCES Working Paper 97-31). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Levine, R., Huberman, M., and Buckner, K. (2002). *The Measurement of Instructional Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth- and Eighth-Grade Students and Teachers to Questionnaire Items* (NCES Working Paper 2002-06). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

McLaughlin, D. (1997). *Can State Assessment Data Be Used to Reduce State NAEP Sample Sizes?* (NCES Working Paper 97-29). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Mullis, V. (2003). *NAEP Validity Studies: Optimizing State NAEP: Issues and Possible Improvements* (NCES Working Paper 2003-09). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Niemi, R.G. (1997). *Innovative Solutions to Intractable Large-Scale Assessment (Problem 2: Background Questionnaires)* (NCES Working Paper 97-32). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Olson, J.F., and Goldstein, A.A. (1997). *The Inclusion of Students With Disabilities and Limited English Proficient Students in Large-Scale Assessments: A Summary of Recent Progress* (NCES 97-482). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Owen, E.H. (1992). *Exploring New Methods for Collecting Students' School-Based Writing* (NCES 92-065). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Patz, R.J., Wilson, M., and Hoskens, M. (1997). *Optimal Rating Procedures and Methodology for NAEP Open-Ended Items* (NCES Working Paper 97-37). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Pearson, D.P., and Garavaglia, D.R. (2003). *NAEP Validity Studies: Improving the Information Value of Performance Items in Large-Scale Assessments* (NCES Working Paper 2003-08). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Stancavage, F.B. (2003) *NAEP Validity Studies: An Agenda for NAEP Validity Research Validity Research* (NCES Working Paper 2003-07). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Weston, T.J. (2003). *NAEP Validity Studies: The Validity of Oral Accommodation in Testing* (NCES Working Paper 2003-06). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Data Quality and Comparability

Bandeira de Mello, V., Blankenship, C., and McLaughlin, D.H. (2009). *Mapping State Proficiency Standards Onto NAEP Scales: 2005-2007* (NCES 2010-456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Binkley, M., and Kelly, D. (2003). *A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments* (NCES Working Paper 2003-10). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

DeVito, P.J., and Koenig, J.A. (2001). *NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting.* Washington, DC: National Research Council.

Hoffman, G.R., Becker, D.E., and Wise, L. (2003). *NAEP Quality Assurance Checks of the 2002 Reading Assessment Results of Delaware* (NCES Working Paper 2003-19). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Ingels, S., and Taylor, J. (1995). *National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data* (NCES Working Paper 95-06). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Johnson, E.G. (1998). *Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): A Technical Report* (NCES 98-499). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Johnson, E., Cohen, J., Chen, W.H., Jiang, T., and Zhang, Y. (2005). *2000 NAEP–1999 TIMSS Linking Report* (NCES Working Paper 2005-01). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Kitmitto, S., and Bandeira de Mello, V., (2008). *Measuring the Status and Change of NAEP State Inclusion Rates for Students with Disabilities* (NCES 2009–453). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Neidorf, T.S., Binkley, M., Gattis, K., and Nohara, D. (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006-029). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Neidorf, T.S., Binkley, M., and Stephens, M. (2006). *Comparing Science Content in the National Assessment of Educational Progress (NAEP) 2000 and Trends in International Mathematics and Science Study (TIMSS) 2003 Assessments* (NCES 2006-026). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Pellegrino, J.W., Jones, L.R., and Mitchell, K.J. (1999). *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress*. Washington, DC: National Research Council.

Phillips, G. W (2009), *The Second Derivative: International Benchmarks in Mathematics for U.S. States and School Districts*. Washington, DC: American Institutes for Research.

Raju, N.S., Pellegrino, J.W., Bertenthal, M.W., Mitchell, K.J., and Jones, L.R. (2000). *Grading the Nation's Report Card: Research From the Evaluation of NAEP*. Washington, DC: National Research Council.

Sedlacek, D.A. (1995). *Model-Based Methods for Analysis of Data From 1990 NAEP Trial State Assessment* (NCES 95-696). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Sedlacek, D.A. (1995). *Use of Person-Fit Statistics in Reporting and Analyzing National Assessment of Educational Progress Results* (NCES 95-713). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Sedlacek, D.A. (1995). *Using HLM and NAEP Data to Explore School Correlates of 1990 Mathematics and Geometry Achievement in Grades 4, 8, 12—Methodology and Results* (NCES 95-697). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Chapter 19: National Adult Literacy Survey (NALS)

1. OVERVIEW

PERIODIC SURVEY OF A SAMPLE OF ADULTS LIVING IN HOUSEHOLDS OR PRISONS

Assesses literacy skills:

- Prose
- Document
- Quantitative

Collects background data:

- Demographics
- Education
- Labor market experiences
- Income
- Activities

The National Adult Literacy Survey (NALS) was initiated to fill the need for accurate and detailed information on the English literacy skills of America's adults. In accordance with a congressional mandate, it provided the most detailed portrait that has ever been available in the 1990s on the condition of literacy in this nation.

The 1992 NALS is the third assessment of adult literacy funded by the federal government and conducted by the Educational Testing Service (ETS). The two previous efforts were (1) the 1985 Young Adult Literacy Assessment, funded as an adjunct to the National Assessment of Educational Progress (NAEP)—see chapter 18); and (2) the Department of Labor's 1990 Workplace Literacy Survey. Building on these two earlier surveys, literacy for NALS is defined along three dimensions—prose, document, and quantitative—designed to capture an ordered set of information-processing skills and strategies that adults use to accomplish a diverse range of literacy tasks encountered in everyday life. The background data collected in NALS provide a context for understanding the ways in which various characteristics are associated with demonstrated literacy skills.

NALS is the first national study of literacy for *all* adults since the Adult Performance Level Surveys conducted in the early 1970s. It is also the first in-person literacy assessment involving the prison population. A second adult literacy survey, the National Assessment of Adult Literacy (NAAL), was conducted in 2003.

Purpose

To (1) evaluate the English language literacy skills of adults (16 years and older) living in households or prisons in the United States; (2) relate the literacy skills of the nation's adults to a variety of demographic characteristics and explanatory variables; and (3) compare the results with those from the 1985 Young Adult Literacy Assessment and the 1990 Workplace Literacy Survey.

Components

The 1992 survey consisted of one component that was administered to three different representative samples: a national household sample; supplemental state household samples for 12 states (California, Florida, Illinois, Indiana, Iowa, Louisiana, New Jersey, New York, Ohio, Pennsylvania, Texas, and Washington); and a national sample of federal and state prison inmates. Responses from the national, state, and prison samples were combined to yield the best possible performance estimates.

National Adult Literacy Survey. The 1992 survey assessed the literacy skills of a representative sample of the U.S. adult population using simulations of three kinds of literacy tasks that adults would ordinarily encounter in daily life(prose, document, and quantitative literacy). The data were collected through in-person

interviews with adults who were living in households or in federal or state prisons. Adults were defined as individuals 16 years or older for the national and prison samples, and 16 to 64 years of age for the state samples. In addition to the cognitive tasks, the personal interview gathered information on demographic characteristics, language background, educational background, reading practices, and labor market experiences. To ensure comparability across all samples, the literacy tasks assessed were the same for all three samples. Background data varied somewhat between the household and prison samples—labor force questions were irrelevant to prisoners, and questions about criminal behavior and sentences were relevant only to prisoners.

Literacy Assessment. The pool of literacy tasks used to measure adult proficiencies consisted of 165 literacy questions—41 prose, 81 document, and 43 quantitative. To ensure that valid comparisons could be made by linking the scales to those of the 1985 Young Adult Literacy Assessment, 85 tasks from that survey were included in the 1992 survey. An additional 80 new tasks were developed specifically to complement and enhance the original 85 tasks. The literacy tasks administered in NALS varied widely in terms of materials and content. The six major context/content areas were home and family; health and safety; community and citizenship; consumer electronics; work; and leisure and recreation. Each adult was given a subset (about 45) of the total pool of assessment tasks to complete. Each of the tasks extended over a range of difficulty on the three literacy scales. The new tasks were designed to simulate the way in which people use various types of materials and to require different strategies for successful performance.

The responses to the literacy assessment were pooled and reported by proficiency scores, ranging from 0 to 500, on three separate scales, one each for prose, document, and quantitative literacy. By examining the overall characteristics of individuals who performed at each literacy level on each scale, it is possible to identify factors associated with higher or lower proficiency in reading and using prose, document, and quantitative materials.

Background Information. Background information collected for the state and household samples included data on *background and demographics*—country of birth, languages spoken or read, access to reading materials, size of household, educational attainment of parents, age, race/ethnicity, and marital status; *education*—highest grade completed in school, current aspirations, participation in adult education classes, and education received outside the country; *labor market experiences*—employment status, recent labor market experiences, and occupation; *income*—personal and household; and *activities*—voting behavior, hours spent watching television, frequency and content of newspaper reading, and use of literacy skills for work and leisure. Respondents from each of the 12 participating states were also asked state-specific questions.

To address issues of particular relevance to the prison population, a separate background questionnaire was developed for the prison sample. This instrument drew questions from the 1991 Survey of Inmates of State Correctional Facilities, sponsored by the Department of Justice's Bureau of Justice Statistics. The background questionnaire for the prison population addressed the following major topics: general and language background; educational background and experience; current offenses and criminal history; prison work assignments and labor force participation prior to incarceration; literacy activities and collaboration; and demographic information.

Periodicity

NALS was conducted in 1992. NAAL, a continuation of NALS, was conducted in 2003.

2. USES OF DATA

Results from NALS provide a detailed portrait on the condition of literacy in this nation. NALS data provide vital information to policymakers, business and labor leaders, researchers, and citizens. The survey results can be used to

- describe the levels of literacy demonstrated by the adult population as a whole and by adults in various subgroups (e.g., those targeted as at risk, prison inmates, and older adults);
- characterize adults' literacy skills in terms of demographic and background information (e.g., reading characteristics, education, and employment experiences);
- profile the literacy skills of the nation's workforce;
- compare assessment results from the current study with those from the 1985 Young Adult Literacy Assessment;
- interpret the findings in light of information-processing skills and strategies, so as to inform

curriculum decisions concerning adult education and training; and

- increase understanding of the skills and knowledge associated with living in a technological society.

3. KEY CONCEPTS

Some of the key concepts related to the literacy assessment are described below. See the NALS Electronic Codebook or appendices of NALS reports for lists and descriptions of variables.

Literacy. The ability to use printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential. This definition goes beyond simply decoding and comprehending text to include a broad range of information-processing skills that adults use in accomplishing the range of tasks associated with work, home, and community contexts.

Prose Literacy. The ability to locate information contained in expository or narrative prose in the presence of related but unnecessary information, find all of the relevant information, integrate information from various parts of a passage of text, and write new information related to the text. Expository prose consists of printed information in the form of connected sentences and longer passages that define, describe, or inform, such as newspaper stories or written instructions. Narrative prose tells a story, but is less frequently used by adults in everyday life than by school children, and did not occur as often in the text presented in NALS as prose literacy tasks. Prose varies in its length, density, and structure.

Document Literacy. The ability to locate information in documents, repeat the search as many times as needed to find all the information, integrate information from various parts of a document, and write new information as requested in appropriate places in a document, while screening out related but inappropriate information. Documents differ from prose text in that they are more highly structured. Documents consist of structured prose and quantitative information in complex arrays arranged in rows and columns, such as tables, data forms, and lists (simple, nested, intersected, or combined); in hierarchical structures, such as tables of contents or indexes; or in two-dimensional visual displays of quantitative information, such as graphs, charts, and maps.

Quantitative Literacy. The ability to use quantitative information contained in prose or documents (specifically the ability to locate quantities while screening out related but unneeded information), repeat the search as many times as needed to find all the numbers, integrate information from various parts of a text or document, infer the necessary arithmetic operation(s), and perform arithmetic operation(s). Quantities can be located in either prose texts or in documents. Quantitative information may be displayed visually in graphs, maps, or charts, or it may be displayed numerically using whole numbers, fractions, decimals, percentages, or time units (hours and minutes).

Literacy Scales. Three scales used to report the results for prose, document, and quantitative literacy. These scales, each ranging from 0 to 500, are based on those established for the 1985 Young Adult Literacy Assessment. The scores on each scale represent degrees of proficiency along that particular dimension of literacy. The literacy tasks administered in the 1992 survey varied widely in terms of materials, content, and task requirements, and thus in difficulty. A careful analysis of the range of tasks along each scale provides clear evidence of an ordered set of information-processing skills and strategies along each scale. To capture this ordering, each scale was divided into five levels that reflect this progression of information-processing skills and strategies: Level 1 (0 to 225), Level 2 (226 to 275), Level 3 (276 to 325), Level 4 (326 to 375), and Level 5 (376 to 500). Level 1 comprised those adults who could consistently succeed with Level 1 literacy tasks but not with Level 2 tasks, as well as those who could not consistently succeed with Level 1 tasks and those who were not literate enough in English to take the test at all. Adults in Levels 2 through 4 were consistently able to succeed with tasks at their level but not with the next more difficult level of tasks. Adults in Level 5 were consistently able to succeed with Level 5 tasks.

Succeed Consistently. Indicates that a person at or above a given level of literacy has at least an 80 percent chance of correctly responding to a particular task. This 80 percent criterion is more stringent than the 65 percent standard used in NAEP (see chapter 18) for measuring what school children know and can do.

4. SURVEY DESIGN

The 1992 NALS was designed and administered by ETS. A subcontract was awarded to Westat, Inc., for sampling and field data collection. A committee of

experts from business and industry, labor, government, research, and adult education worked with the ETS staff to develop the definition of literacy that underlies NALS, as well as to prepare the assessment objectives that guided the selection and construction of assessment tasks. In addition to this Literacy Definition Committee, a Technical Review Committee was formed to help ensure the soundness of the assessment design, the quality of the data collected, the integrity of the analyses conducted, and the appropriateness of the interpretations of the final results. The prison survey was developed in consultation with the Bureau of Justice Statistics and the Federal Bureau of Prisons. The survey design for the 1992 survey is described below.

Target Population

The target population for the national household sample consisted of adults 16 years and older in the 50 states and the District of Columbia who, at the time of the survey, resided in private households or college dormitories. The target population for the supplemental state household sample consisted of individuals 16 to 64 years of age who, at the time of the survey, resided in private households or college dormitories in the participating state (California, Florida, Illinois, Indiana, Iowa, Louisiana, New Jersey, New York, Ohio, Pennsylvania, Texas, or Washington). Individuals residing in other institutions—nursing homes, group homes, or psychiatric facilities—were not included in the household samples. The target population for the prison sample consisted of adults 16 years or older who were in state or federal prisons at the time of the survey; those held in local jails, community-based facilities, or other types of institutions were not included.

Sample Design

Because this 1992 survey was designed to provide data representative at the national level (including prison inmates) and at the state level for participating states, it included three different samples: a national household sample, supplemental state household samples for 12 states, and a supplemental national sample of state and federal prison inmates.

Household Samples. The sample design for the national and state household samples involved a four-stage stratified area sample: (1) the selection of primary sampling units (PSUs) consisting of counties or contiguous groups of counties; (2) the selection of segments (within the selected PSUs) consisting of census blocks or groups of contiguous census blocks; (3) the selection of households within the segmented samples; and (4) the selection of age-eligible individuals within each selected household. The sample design requirements called for an average cluster size of seven interviews (i.e., seven completed background questionnaires per segment). In addition, a reserve sample at the household level of approximately 5 percent of the size of the main sample was selected and set aside in case of shortfalls due to unexpectedly high vacancy and nonresponse rates.

One national area sample was drawn for the national household sample, and 12 independent state-specific area samples were drawn from the 12 states participating in the supplemental state samples. The sample designs used for all 13 samples were similar, with one major difference. In the national sample, Black and Hispanic respondents were sampled at about double the rate of the remainder of the population to assure reliable estimates of their literacy proficiencies, whereas the state samples used no oversampling.

The first stage of sampling involved the selection of PSUs. A national sampling frame of 1,400 PSUs was constructed primarily from 1990 census data stratified on the basis of region, metropolitan status, percent Black, percent Hispanic, and whenever possible, per capita income. Using this frame, 101 PSUs were selected for the national sample. The national frame of PSUs (subdivided at state boundaries, if needed) was used to construct individual state frames for the supplemental state sample; a sample of 8 to 12 PSUs was selected within each of the given states. All PSUs were selected with probability proportional to the PSU's 1990 population.

The second stage of sampling involved the selection of segments within the selected PSUs. The Bureau of the Census's Topologically Integrated Geographical Encoding and Referencing (TIGER) System File was used for the production of segment maps. The segments were selected with probability proportional to size, where the measure of size for a segment was a function of the number of year-round housing units within the segment. The oversampling of Black and Hispanic respondents for the national sample was carried out at the segment level, where segments were classified either as having a high percentage of the Black or Hispanic population (more than 25 percent) or as not having a high percentage.

The third stage of sampling involved the selection of households within the segmented samples. Westat field staff visited all selected segments in the fall of 1991 and prepared lists of all housing units within the boundaries of each segment as determined by the 1990 census block maps. The lists were used to construct the sampling frame for households. Households were selected with equal probability within each segment,

except for White, non-Hispanic households in segments with a high percentage of the Black or Hispanic population (over 25 percent) in the national sample, which were subsampled so that the sampling rates for White, non-Hispanic respondents would be about the same overall.

The fourth stage of sampling involved the selection of one or two adults within each selected household during the data collection phase of the survey. One person was selected at random from households with fewer than four eligible members; two persons were selected from households with four or more eligible members. Using a screener, the interviewer constructed a list of age-eligible household members (16 and older for the national sample, 16 to 64 for the state sample) for each selected household. The interviewers, who were instructed to list the eligible household members in descending order by age, then identified one or two household members to interview, based on computer-generated sampling messages that were attached to each questionnaire in advance.

Prison Sample. There were two stages of selection for the prison sample. The first stage involved the selection of state or federal correctional facilities. The sampling frame for the correctional facilities was based on the 1990 census of federal and state prisons, updated in mid-1991. The facility frame was stratified prior to sample selection on the basis of type of facility (federal or state prison), region of country, inmate gender composition, and type of security. A sample of 88 facilities and a reserve sample of 8 facilities was then drawn from the frame based on probability proportional to size, where the measure of size for a given facility was equal to the inmate population. The second stage of sampling involved the selection of inmates within each selected facility, using a list of names obtained from the facility administrators. An average of 12 inmates were selected from each facility based on a probability inversely proportional to their facility's inmate population (up to a maximum of 22 interviews in a facility), so that the product of the first- and second-stage probabilities would be constant.

Assessment Design

Building on the 1985 Young Adult Literacy Assessment and the 1991 Workplace Literacy Survey, the NALS Technical Committee adopted the definition of literacy and the literacy scales—prose, document, and quantitative—used in the previous surveys. The materials were selected to represent a variety of contexts and contents: home and family; health and safety; community and citizenship; consumer electronics; work; and leisure and recreation.

BIB Spiraling. The survey design gave each respondent a subset of the total pool of literacy tasks, while at the same time ensuring that each of the 165 tasks was administered to a nationally representative sample of the adult population. The design most suitable for this purpose is a variant of standard matrix sampling called balanced incomplete block (BIB) design.

Literacy tasks were assigned to blocks or sections that could be completed in about 15 minutes, and these blocks were then compiled into booklets so that each block appeared in each position (first, middle, and last) and each block was paired with every other block. Thirteen blocks of simulation tasks were assembled into 26 unique booklets, each of which contained four blocks of tasks: the core (the same for all exercise booklets) and three cognitive blocks. Each booklet could be completed in about 45 minutes.

Pretests. A field test of the national household sample was conducted in the spring of 1991 using a sample of 2,000 adults drawn from 16 PSUs. The purposes of the field test were to evaluate the impact of incentives on response rates, performance, and survey costs; to evaluate newly developed literacy exercises for item bias and testing time; and to evaluate the administration and appropriateness of the background questions. As a result of the field test, some of the literacy tasks and their scoring guides were revised or dropped from the final assessment.

For the prison sample, a small pretest was conducted at the Roxbury Correctional Institution in Hagerstown, Maryland. This pretest was designed to evaluate the ease of administration of the survey instruments, survey administration time, within-facility procedures, and inmate reaction to the survey. The pretest demonstrated that several changes to the background questionnaire would facilitate administration. Administrative procedures were also refined to reflect lessons learned during the pretest.

Data Collection and Processing

The survey data were collected through in-person household or prison interviews during the first 8 months of 1992. As field operations were completed, the data were shipped to ETS for processing. Further description follows.

Reference Dates. Respondents answered the employment status and weekly wages questions for the week before the survey was administered.

Data Collection. During January and February of 1992, field interviewers, supervisors, and editors received

extensive training both in general and survey-specific interview techniques. The NALS field period began in February 1992, immediately following the completion of the first interviewer training sessions, and lasted 28 weeks, until the end of August. All three survey sample groups were worked simultaneously (except for the state of Florida, where data were not collected until 1993). Except for a small, experimental "no incentive" group, all household participants who completed as much of the assessment as their skills allowed received $20 for their time. More than 400 trained interviewers visited about 44,000 households to select and interview almost 31,000 adults. In addition, over 1,147 prison inmates at 87 facilities were interviewed.

Each survey participant was asked to spend approximately one hour responding to survey questions and tasks. Data collection instruments included the screener (designed to enumerate household members and select survey respondents), the background questionnaire, and the literacy exercise booklets. Answering the screener and background questionnaire required no reading or writing skills; to ensure standardized administration, the questions on each were read to respondents in English or Spanish and the answers recorded by the assessment interviewer. Each of the exercise booklets had a corresponding interview guide, with specific instructions to the interviewer for directing the exercise booklet. Reading and writing skills in the English language were required to complete the exercise booklet. When a sampled respondent did not complete any or all of the survey instruments, the interviewer was required to complete a noninterview report form. Field supervisors reviewed the noninterview forms to determine the case's potential for conversion, and the data collected on the form were processed for nonresponse analysis.

Following the completion of an interview, interviewers edited all materials for legibility and completeness. The interviewers sent their completed work to their regional supervisors for a complete edit of the instruments, quality control procedures, and any required data retrieval. As these tasks were completed, the cases were shipped to ETS for processing.

During the data collection process, two special quality control procedures were implemented to identify any households or dwellings missed during the listing phase: the missing structure procedure and the missed dwelling unit procedure. These procedures were used to give these missed structures and dwelling units a chance of selection at time of data collection.

The field effort occurred in three overlapping stages:

(1) *Initial Phase.* Each area segment was assigned by the regional supervisor to an interviewer, who followed certain rules in making a prescribed number of calls (a maximum of four was used) to every sampled dwelling in the segment.

(2) *Reassignment Phase.* Cases that did not result in completed interviews during the initial phase were reviewed by the regional supervisor, and a subset was selected for reassignment to another interviewer in the same PSU or an interviewer from a nearby PSU.

(3) *Special Nonresponse Conversion Phase.* The home office assembled a special traveling team of the most experienced or productive interviewers to perform a nonresponse conversion effort, under the supervision of a subset of the field supervisors.

Data Processing. Coding and scoring staff underwent intensive training prior to the actual coding and scoring. A scoring supervisor monitored both the coding of the questionnaires and the scoring of the exercise booklets. The background questionnaire was designed to be read by a computerized scanning device. Nearly all the simulation tasks contained in the exercise booklet were open-ended; with scoring guides as examples, responses to these items were classified as correct, incorrect, or omitted by trained readers. Responses from the screener and scores from the exercise booklets were transferred to scannable answer sheets. Each survey instrument's scannable forms were batched and sent to the scanning department at regular intervals. As the different instruments were processed, the data were transferred to a database on the main ETS computer for editing.

Editing. Several quality control procedures related to data collection were used during the field operation: an interviewer field edit, a complete edit of all documents by a trained field editor, validation of 10 percent of each interviewer's closeout work, and field observation of both supervisors and interviewers. Additional edits were done during data processing. These included an assessment of the internal logic and consistency of the data received. Discrepancies were corrected whenever possible. The background questionnaires were also checked to make sure that the skip patterns had been followed and all data errors were resolved. In addition, a random set of exercise booklets was selected to provide an additional check on the accuracy of transferring information from booklets and answer sheets to the database.

Estimation Methods

Weighting was used in the 1992 NALS, prior to the calculation of base weights. Responses to the literacy tasks were scored using item response theory (IRT) scaling. A multiple imputation procedure based on plausible values methodology was used to estimate the literacy proficiencies of individuals who completed literacy tasks. An innovative approach was implemented to impute missing cognitive data in order to minimize distortions in the population proficiency estimates due to nonresponse to the literacy booklet.

Weighting. Full sample and replicate weights were calculated for survey respondents who completed the exercise booklet; those who could not start the exercises because of a language barrier, a physical or mental barrier, or a reading or writing barrier; and those who refused to complete the exercises but had completed background questionnaires. Demographic variables critical to the weighting were recoded and imputed, if necessary, prior to the calculation of base weights (see "Imputation" below). Separate sets of weights were computed for the incentive and "no incentive" samples.

Household samples. A base weight was computed for each eligible record. The base weight initially was computed as the reciprocal of the product of probabilities of selection for a respondent at the PSU, segment, dwelling unit, and person levels. The final base weight included adjustments to reflect the selection of the reserve sample, the selection of missed dwelling units, and the chunking process conducted during the listing of the segments; and to account for the subsample of segments assigned to the "no incentive" experiment and the subsampling of respondents within households. The base weights for each sample were then poststratified to known 1990 census population totals, adjusted for undercount. This first-level stratification provided sampling weights with lower variation and adjusted for nonresponse. State records were poststratified separately from national records to provide a common base for applying composite weighting factors; population totals were calculated separately for each distinct group.

Composite weights were developed so that NALS data could be used to produce both state and national statistics. For the household samples, a composite weight was computed as the product of the poststratified base weight and a compositing factor that combined the national and state sample data in an optimal manner, considering the differences in sample design, sample size, and sampling error between the two sampled groups. Up to four different compositing factors were used in each of the 11 participating states, and a pseudo-factor (equal to 1) was used for all persons 65 and older and for all national sample records from outside the 11 participating states.

To compute the final sample weights, the composite weights were adjusted to known 1990 census counts (adjusted for undercount), using a process called the poststratification raking ratio adjustment. The cells used for raking were defined to the finest combination of age, race/ethnicity, sex, education, and geographic indicators (e.g., Metropolitan Statistical Area [MSA] vs. non-MSA) that the data would allow. Raking adjustment factors were calculated separately for each of the state samples and then for the remainder of the United States.

The above steps used to create the final sample weights were repeated for 60 strategically constructed subsets of the household sample to create a set of replicate weights to be used for variance estimation using the jackknife method.

Prison sample. Base weights for the prison respondents were constructed to be equal to the reciprocal of the product of the selection probabilities for the facility and the inmate within the facility. These weights were then nonresponse-adjusted to reflect both facility and inmate nonresponse. To compute the final sample weights, the resulting nonresponse-adjusted weights were then raked to agree with independent estimates for certain subgroups of the prison population. The above procedures were repeated for 45 strategically constructed subsets of the prison sample to create a set of replicate weights to be used for variance estimation using the jackknife method.

Scaling. Since NALS used a variant of matrix sampling and since different respondents received different sets of tasks, it would be inappropriate to report its results using conventional scoring methods based on the number of correct responses. The literacy assessment results are reported using IRT scaling, which assumes some uniformity in response patterns when items require similar skills. Such uniformity can be used to characterize both examinees and items in terms of a common scale attached to the skills, even when all examinees do not take identical sets of items. Comparisons of items and examinees can then be made in reference to a scale, rather than to the percent correct. IRT scaling also allows the distributions of examinee groups to be compared.

The results of the 1992 literacy assessment are reported on three scales (prose, document, and quantitative) that were established for the 1985 Young Adult Literacy Assessment. Separate IRT linking and scaling were

carried out for each of the three domains, using the three-parameter logistic (3PL) scaling model from item response theory. This is a mathematical model for estimating the probability that a particular person will respond correctly to a particular item from a single domain of items. The probability is given as a function of a parameter characterizing the proficiency of that person and three parameters characterizing the properties of that item. Item parameters needed for the 3PL scaling model were estimated by linking each of the literacy scales used in the 1992 survey to the 1985 Young Adult Literacy Assessment scales.

Imputation. Imputation was performed prior to weighting on missing demographic items considered critical to weighting. Literacy proficiencies of respondents were estimated using a multiple imputation procedure based on plausible values methodology. Missing cognitive data were also imputed.

Demographic data. Demographic variables critical to the weighting (race/ethnicity of the head of household; sex, age, race/ethnicity, and education of the respondent) were recoded and collapsed to required levels, and imputed, if necessary, prior to the calculation of base weights. Data from the background questionnaire were preferred for all items except race/ethnicity of the head of household, which was collected in the screener. For the few cases in which the background questionnaire measure was missing, the screener measure was generally available and was used as a direct substitute. The amount of missing data remaining after substitution was small, making the imputation task fairly straightforward. A standard (random within class) hot-deck imputation procedure was performed for particular combinations of fields that were missing. Imputation flags were created for each of the five critical fields to indicate whether data were originally reported or were based on substitution or imputation. The imputed values were used only for the sample weighting process.

Literacy proficiency estimation (plausible values). A multiple imputation procedure based on plausible values methodology was used to estimate respondents' literacy proficiency in the 1992 NALS. When analyzing the distribution of proficiencies in a group of persons, more efficient estimates can be obtained from a sample design similar to that used in this 1992 survey. Such designs solicit relatively few cognitive responses from each sampled respondent, but maintain a wide range of content representation when responses are summed for all respondents.

In the 1992 survey, all proficiency data were based on two types of information: responses to the background questions and responses to the cognitive items. As an intermediate step, a functional relationship between the two sets of information was calculated for the total sample, and this function was used to obtain unbiased proficiency estimates for population groups with reduced error variance. Possible values for a respondent's proficiency were sampled from a posterior distribution that is the product of two functions: the conditional distribution of proficiency given the pattern of background variables and the likelihood function of proficiency given the pattern of responses to the cognitive items. Since exact matches of background responses are quite rare, NALS used more than 200 principal components to summarize the background information, capturing more than 99 percent of the variance. More detailed information on the plausible values methodology used in the 1992 survey is available in the *Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey* (Kirsch et al. 2000).

Cognitive data. New procedures were implemented in the 1992 NALS to minimize distortions in the population proficiency estimates due to nonresponse to the literacy booklets. When a sampled individual decided to stop the assessment (answered less than five literacy items per scale), the interviewer used a standardized nonresponse coding procedure to record the reason why the person was stopping. This information was used to classify nonrespondents into two groups: (1) those who stopped the assessment for literacy-related reasons (e.g., language difficulty, mental disability, or reading difficulty not related to a physical disability); and (2) those who stopped for reasons unrelated to literacy (e.g., physical disability or refusal). About half of the individuals did not complete the assessment for reasons related to their literacy skills; the other respondents gave no reason for stopping or gave reasons unrelated to their literacy.

To represent the range of implied causes of missing literacy responses, the imputation procedure selected relied on background variables and self-reported reasons for nonresponse, in addition to the functional relationship between background variables and proficiency scores for the total population. It treated "consecutively missing" data from the literacy booklet instrument differently depending on whether the nonrespondents' reasons were related or unrelated to their literacy skills: (1) those who gave literacy-related reasons were treated as wrong answers, based on the assumption that they could not have correctly completed the literacy tasks, whereas (2) those who gave no reason or cited reasons unrelated to literacy

skills for not completing the assessment were essentially ignored (considered not reached), since it could not be assumed that their answers would have been either correct or incorrect. The proficiencies of such respondents were inferred from the proficiencies of other adults with similar characteristics using the plausible values methodology described above.

Future Plans

A second survey, NAAL, was conducted in 2003. Currently, there are no plans to administer another measure of adult literacy.

5. DATA QUALITY AND COMPARABILITY

The NALS sampling design and weighting procedures assured that participants' responses could be generalized to the population of interest. In addition, NCES conducted special evaluation studies to examine issues related to the quality of NALS. These studies included (1) a study of the role of incentives in literacy survey research; (2) an evaluation of its sample design and composite estimation; and (3) an evaluation of the construct validity of the adult literacy scales.

Sampling Error

In the 1992 survey, the use of a complex sample design, adjustments for nonresponse, and poststratification procedures resulted in dependence among the observations. Therefore, a jackknife replication method was used to estimate the sampling variance. The mean square error of replicate estimates around their corresponding full sample estimate provides an estimate of the sampling variance of the statistic of interest. The replication scheme was designed to produce stable estimates of standard errors for national and prison estimates as well as for the 12 individual states.

The advantage of compositing the national and state samples during sample weighting was the increased sample size, which improved the precision of both the state and national estimates. However, biases could be present because the national PSU sample strata were not designed to maximize the efficiency of state-level estimates.

Nonsampling Error

The major source of nonsampling error in the 1992 NALS was nonresponse error; special procedures were developed to minimize potential nonresponse bias based on how much of the survey the respondent completed. Other possible sources of nonsampling error were random measurement error and systematic error due to interviewers, coders, or scorers.

Coverage Error. Coverage error could result from either the sampling frame of households or prisons being incomplete or from a household's or prison's failure to include all adults 16 years and older on the lists from which the sampled respondents were drawn. Special procedures and edits were built into NALS to review both listers' and interviewers' ongoing work and to give any missed structures and/or dwelling units a chance of selection at data collection. However, just as all other household personal interview surveys have persistent undercoverage problems, the 1992 survey had problems in population coverage due to interviewers not gaining access to households in dangerous neighborhoods, locked residential apartment buildings, and gated communities.

Nonresponse Error.

Unit nonresponse. Since three survey instruments—screener, background questionnaire, and exercise booklet—were required for the administration of the survey, it was possible for a household or respondent to refuse to participate at the time of the administration of any one of these instruments. Because the screener and background questionnaire were read to the survey participants in English or Spanish, but the exercise booklet required reading and writing in the English language, it was possible to complete the screener or background questionnaire but not the exercise booklet, and vice versa. Thus, response rates were calculated for each of the three instruments for the household samples (see table 12). For the prison sample, there were only two points at which a respondent could not respond—at the administration of the background questionnaire or the exercise booklet.

The response rate to the background questionnaire was 80.5 percent. For the household samples, the response rates exclude individuals who were not paid incentives. Also excluded are the respondents to the Florida state survey, which had a delayed administration.

The combined national and state household target sample in the 1992 NALS included 43,780 representative housing units, of which 5,410 were vacant. Approximately 89 percent of the occupied households completed a screener.

The household sample screening effort identified a total of 30,810 eligible respondents, of whom 24,940 (81.0 percent unweighted) completed the background questionnaire. For the prison sample, 87 of the 88 sampled facilities participated in the survey. Of the 1,340 inmates selected, 1,150 (85.6 percent

unweighted) completed the background questionnaire. For the occupied households, "refusal or breakoff" was the most common explanation for nonresponse to the screener and background questionnaire. The second most common explanation was "not at home after maximum number of calls." Nonresponse also resulted from language, physical, and mental problems. Housing units or individuals who refused to participate before any information was collected about them, or who did not answer a sufficient number of background questions, were never incorporated into the database. Because these individuals were unlikely to know that the survey intended to assess their literacy, it was assumed that their reason for not completing the survey was not related to their level of literacy.

Literacy assessment booklets were considered complete if at least five items were answered on each scale. A total of 24,940 household sample members were classified as eligible for the exercise booklet. Of these, 88.6 percent completed the booklet and another 6.1 percent partially completed it. Of the 1,150 eligibles in the prison sample, 86.8 percent completed the booklet and another 9.3 percent partially completed it.

There were reasons to believe that the literacy performance data were missing more often for adults with lower levels of literacy than for adults with higher levels. Field-test evidence and experience with surveys indicated that adults with lower levels of literacy were more likely than adults with higher proficiencies either to decline to respond to the survey at all or to begin the assessment but not complete it. Ignoring this pattern of missing data would have resulted in overestimating the literacy skills of adults in the United States. Therefore, to minimize bias in the proficiency estimates due to nonresponse to the literacy assessment, special procedures were developed to impute the literacy proficiencies of nonrespondents who completed fewer than five literacy tasks.

Item nonresponse. For each background questionnaire, staff verified that certain questions providing critical information for weighting and data analyses had been answered, namely, education level, employment status, parents' level of education, race, and sex. If a response was missing, the case was returned to the field for data retrieval. Therefore, item response rates for completed background questionnaires were quite high, although they varied by type of question. Questions asking country of origin (first question in the booklet) and sex (last question in the booklet) had nearly 100 percent response rates, indicating that most respondents attempted to complete the entire questionnaire. Response rates were lower, however, for questions about income and educational background.

Table 12. Weighted and unweighted response rates for all sample types in the National Adult Literacy Survey, by survey component: 1992

Component	Weighted (percent)	Unweighted (percent)
Screener	—	89.1
Background questionnaire	80.5	81.0
Exercise booklet	95.9	95.9

— Not available.
NOTE: The weighted response rates were calculated by applying the sampling weight to each individual to account for his or her probability of selection into the sample. Weighted response rates were computed only for screened households (the probability of selection is not known for persons in households that were not screened).
SOURCE: Kirsch, I.S., Yamamoto, K., Norris, N., Rock, D., Jungeblut, A., O'Reilly, P., Campbell, A., Jenkins, L., Kolstad, A., Berlin, M., Mohadjer, L., Waksberg, J., Goksel, H., Burke, J., Rieger, S., Green, J., Klein, M., Mosenthal, P., and Baldi, S. (2000). *Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey* (NCES 2001-457). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

The electronic codebook provides counts of item nonresponse. These, however, have to be considered in terms of the number of adults that were offered each task, because a great deal of the missing data is missing by design.

Measurement Error. All background questions and literacy tasks underwent extensive review by subject area and measurement specialists, as well as scrutiny to eliminate any bias or lack of sensitivity to particular groups. Special care was taken to include materials and tasks that were relevant to adults of widely varying ages. During the test development stage, the tasks were submitted to test specialists for review, part of which involved checking the accuracy and completeness of the scoring guide. After preliminary versions of the assessment instruments were developed and after the field test was conducted, the literacy tasks were closely analyzed for bias or "differential item functioning." The goal was to identify any assessment tasks that were likely to underestimate the proficiencies of a particular subpopulation, whether it be older adults, females, or Black or Hispanic adults. Any assessment item that appeared to be biased against a subgroup was excluded from the final survey. The coding and scoring guides also underwent further revisions after the first responses were received from the main data collection.

Interviewer error checks. Several quality control procedures related to data collection were used during the field operation: an interviewer field edit, a complete edit of all documents by a trained field editor, validation of 10 percent of each interviewer's closeout work, and field observation of both supervisors and interviewers.

Coding/scoring error checks. In order to monitor the accuracy of coding, the questions dealing with country of birth, language, wages, and date of birth were checked in 10 percent of the questionnaires by a second coder. For the industry and occupation questions, 100 percent of the questionnaires were recoded by a second coder. Twenty percent of all the exercise booklets were subjected to a reader reliability check, which entailed a scoring by a second reader. There was a high degree of reader reliability across tasks—ranging from 88.1 to 99.9 percent—with an average agreement of 97 percent. For 133 out of 165 open-ended tasks, the agreement between the two readers was above 95 percent.

Data Comparability

One of the major goals of this survey was to compare its results to the 1985 Young Adult Literacy Assessment and other large assessment studies. NALS is also comparable with NAAL, conducted in 2003, in terms of assessment scores (see chapter 20).

Comparisons with the 1985 Young Adult Literacy Assessment. Comparisons are possible because the sample design, item pool, and methodology used in the 1985 Young Adult Literacy Assessment and the 1992 survey were very similar. Literacy tasks for each survey were developed using the same definition of literacy, and a subset of identical tasks was administered in both assessments. Scoring guides were the same for both surveys. Both gave nearly identical incentive payments to participants ($15 in 1985 and $20 in 1992). The literacy scales used in the two surveys were linked so that the scores could be reported on a common scale.

Nevertheless, there were some differences in procedures for the two surveys. For example, missing responses to the literacy tasks were handled differently. In the 1985 Young Adult Literacy Assessment, individuals who could not answer six core literacy tasks and those who spoke only Spanish were excluded from the analyses. In the 1992 survey, however, a special procedure was used to impute literacy proficiencies for literacy-related nonrespondents.
Due to such procedural differences, direct comparisons of the results of the two surveys are not simple and straight-forward. However, because the 1992 sample is more inclusive than the 1985 sample, subsamples that have more exact counterparts in the 1985 survey can be selected. For instance, the initial report from the 1992 NALS presented data, using no subsample matching that indicated that young adults in 1992 were somewhat less literate than their predecessors in 1985. However, when a comparison was made between matched subsamples of the 1985 and 1992 survey respondents based on reasons for nonresponse, the proficiency differences decreased significantly. Furthermore, results from partition analysis of the two surveys' matched subsamples—based on change due to variations in demographic characteristics versus change not related to demography—suggest that most of the observed declines in the average literacy skills of young adults over time can be accounted for by shifts in the composition of the population and by changes across the assessments in the rules used to include or exclude nonrespondents.

Comparisons with the 1993 General Educational Development (GED) Tests. Comparisons between NALS and GED examinees are explored in *The Literacy Proficiencies of GED Examinees: Results From the GED-NALS Comparison Study* (Baldwin et al. 1993). The GED tests and NALS instruments have a considerable degree of overlap in what they measure. Both assess skills that appear to represent verbal comprehension and reasoning or the ability to understand, analyze, interpret, and evaluate written information and apply fundamental principles and concepts. Despite the considerable degree of overlap, the two instruments also measure somewhat different skills. For example, the GED tests seem to tap unique dimensions of writing mechanics and mathematics, while the adult literacy scales appear to tap unique dimensions of document literacy. In addition, the evidence shows that there are no differences in the average prose, document, or quantitative literacy skills of those adults who terminated their schooling at the high school or GED level.

6. CONTACT INFORMATION

For content information on the National Adult Assessments of Literacy, contact

Andrew J. Kolstad
Phone: (202) 502-7374
E-mail: andrew.kolstad@ed.gov

Mailing Address:
National Center for Education Statistics
Institute of Education Sciences

U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

7. METHODOLOGY AND EVALUATION REPORTS

General

Baldwin, J., Kirsch, I.S., Rock, D., and Yamamoto, K. (1993). *The Literacy Proficiencies of GED Examinees: Results From the GED-NALS Comparison Study*. American Council on Education and Educational Testing Service.

Kirsch, I.S., Jungeblut, A., Jenkins, L., and Kolstad, K. (1993). *Adult Literacy in America: A First Look at the Findings of the National Adult Literacy Survey* (NCES 93-275). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Kirsch, I.S., Yamamoto, K., Norris, N., Rock, D., Jungeblut, A., O'Reilly, P., Campbell, A., Jenkins, L., Kolstad, A., Berlin, M., Mohadjer, L., Waksberg, J., Goksel, H., Burke, J., Rieger, S., Green, J., Klein, M., Mosenthal, P., and Baldi, S. (2000). *Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey* (NCES 2001-457). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Survey Design

Campbell, A., and Kirsch, I.S. (1992). *Assessing Literacy: The Framework for the National Adult Literacy Survey* (NCES 92-113). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Chapter 20: National Assessment of Adult Literacy (NAAL)

1. OVERVIEW

The 2003 National Assessment of Adult Literacy (NAAL) is a nationally representative assessment of English literacy among American adults age 16 and older. Sponsored by the National Center for Education Statistics (NCES), NAAL is the nation's most comprehensive measure of adult literacy since the 1992 National Adult Literacy Survey (NALS).

In 2003, over 19,000 adults participated in the national and state-level assessments, representing the entire population of U.S. adults age 16 and older (in households and prisons) in the 50 states and the District of Columbia. Approximately 1,200 of the participants were inmates of state and federal prisons who were assessed separately in order to provide estimates of literacy for the incarcerated population.

By comparing results from 1992 and 2003, NAAL provides the first indicator in a decade of the nation's progress in adult literacy. NAAL also provides information on adults' literacy performance and related background characteristics to researchers, practitioners, policymakers, and the general public.

SURVEY OF A SAMPLE OF ADULTS LIVING IN HOUSEHOLDS OR PRISONS:

Assesses literacy skills:

- Prose
- Document
- Quantitative

Collects background data on:

- Demographics
- Education
- Labor Market Experiences
- Income
- Activities

Purpose

To (1) evaluate the English language literacy skills of adults (age 16 and older) living in households or prisons in the United States; (2) relate the literacy skills of the nation's adults to a variety of demographic characteristics and explanatory variables; and (3) compare the results with those from the 1992 NALS.

Components

NAAL includes a number of components that capture the breadth of adult literacy in the United States: the Background Questionnaire helps identify the relationships between adult literacy and selected demographic and background characteristics; the Prison Component assesses the literacy skills of adults in federal and state prisons; the State Assessment of Adult Literacy (SAAL) gives statewide estimates of literacy for states participating in the state-level assessment; the Health Literacy Component introduces the first-ever national assessment of adults' ability to use their literacy skills in understanding health-related materials and forms; the Fluency Addition to NAAL (FAN) measures basic reading skills by assessing adults' ability to decode, recognize words, and read with fluency; the Adult Literacy Supplemental Assessment (ALSA) provides information on the ability of the least literate adults to identify letters and numbers and to comprehend simple prose and documents; and the main assessment offers a picture of the general literary (i.e., prose, document and quantitative literary) of the adults who passed the core literary tasks.

Background Questionnaire. The 2003 NAAL Background Questionnaire collected data in a variety of background categories; it obtained valuable background information not collected in the 1992 survey. The questionnaire served three purposes:

- to provide descriptive data on respondents;
- to enhance understanding of the factors that are associated with literacy skills used at home, at work, or in the community; and
- to allow for the reporting of changes over time.

The questionnaire was orally administered to every participant by an interviewer who used a computer-assisted personal interview (CAPI) system. Unlike the 1992 NALS, in which the background questions were read aloud from a printed questionnaire, in 2003, interviewers read the questions from laptop computer screens and entered the responses directly into the computer. CAPI then selected the next question based on responses to prior questions. Because the questions were targeted, a respondent did not answer all of the background questions (i.e., inapplicable questions were skipped). The questionnaire took about 28 minutes to complete.

The background questionnaire used in SAAL was the same as that used in NAAL. However, a separate questionnaire was administered for the prison component in order to address issues of particular relevance to the prison population.

Prison Component. The 2003 NAAL Prison component assesses the literacy skills and proficiencies of the U.S. adult prison population. In the 2003 assessment, approximately 1,200 adults participated, from 107 prisons (including 12 federal prisons) in 31 states.

Key features:

- provides demographic and performance data for the prison population, in comparison with the main NAAL household study of the general adult population;
- reports results that are useful to policymakers and practitioners concerned with literacy and education in correctional settings; and
- guides corrections and education professionals in the development of more effective literacy and adult education programs for prison inmates.

The principal aim of the 2003 NAAL prison component is to provide comprehensive information on the literacy and background of the U.S. adult prison population to policymakers and practitioners in order to enhance adult education in our nation's prisons and improve incarcerated adults' ability to function and achieve their goals in the general society, in the workplace, at home, and in the community–upon their release from prison.

State Assessment of Adult Literacy (SAAL). The SAAL is an assessment of adult literacy within a participating state. Conducted in conjunction with the 2003 NAAL data collection, SAAL collected additional data within the six participating states: Kentucky, Maryland, Massachusetts, Missouri, New York, and Oklahoma.

Key features:

SAAL provides participating states with individually-tailored reports that offer:

- more in-depth analysis of a state's literacy, by augmenting the state's sample with the national sample;
- state and national comparisons;
- expanded background information on population groups;
- state-level scoring for FAN, ALSA, and the Health Literacy Component;
- estimates by demographic and other characteristics of interest; and
- trend data (for New York), because it participated in both the 1992 and 2003 assessments.

Health Literacy Component. The 2003 NAAL is the first large-scale national assessment in the United States to contain a component designed specifically to measure health literacy—the ability to use literacy skills to read and understand written health-related information encountered in everyday life. The Health Literacy Component establishes a baseline against which to measure progress in health literacy in future assessments.

The NAAL health literacy report—*The Health Literacy of America's Adults: Results From the 2003 National Assessment of Adult Literacy* (Kutner et al. 2007)—provides first-hand information on the status

of the health literacy of American adults age 16 and older. Results are reported in terms of the four literacy performance levels—below basic, basic, intermediate, and proficient—with examples of the types of health literacy tasks that adults at each level may be able to perform.

Key features:

- reports on the health literacy skills of target audiences;
- sheds light on the relationship between health literacy and background variables, such as educational attainment, age, race/ethnicity, adults' sources of information about health issues, and health insurance coverage;
- examines how health literacy is related to prose, document, and quantitative literacy;
- provides information that may be useful in the development of effective policies and customized programs that address deficiencies in health literacy skills; and
- guides the development of health information tailored to the strengths and weaknesses of target audiences.

Fluency Addition to NAAL (FAN). FAN examines components of oral reading fluency that the main NAAL does not assess. Using speech-recognition software, FAN measures adults' ability to decode, recognize words, and read with fluency.

Key features:

- establishes a basic reading skills scale;
- identifies, for the first time, the relationship between basic reading skills and selected background characteristics, as well as performance on the main NAAL, Health Literacy Component, and prison component; and
- provides a baseline for measuring future changes in the levels and distribution of oral fluency over time.

Ultimately, FAN can improve our understanding of the skill differences between adults who are able to perform relatively challenging tasks and adults who lack basic reading skills. Such information will prove most useful to researchers, practitioners, and policymakers. For instance, adult education providers can use FAN results to develop and offer instruction and courseware that will better address the skill sets of the least literate adults. Likewise, policymakers can use FAN results to support the creation and improvement of programs serving adults with lower literacy skills.

Adult Literacy Supplemental Assessment (ALSA). Low levels of literacy are likely to limit life chances and may be related to social welfare issues, including poverty, incarceration, and preventive health care. Given this, it has become increasingly important for researchers, policymakers, and practitioners to understand the literacy skills and deficits of the least literate adults.

ALSA is designed to assess the basic reading skills of the least literate adults. The 1992 NALS lacked a similar component. Because the least literate adults were unable to complete the 1992 assessment due to literacy-related complications (e.g., difficulty reading and writing in English; mental or learning disabilities), the 1992 NALS provided little information on these respondents.

Key features:

- enhances our understanding of the basic reading skills of the least-literate adults;
- identifies relationships between ALSA scores and selected background characteristics of adults;
- reports results for appropriate demographic groups (e.g., Black, Hispanic, and other racial/ethnic groups; ESL adults; the prison population);
- describes relationships between the performance of ALSA participants and main NAAL participants on the FAN oral reading tasks; and
- provides a baseline for measuring future changes in the levels and distribution of the least literate adults' basic reading skills over time.

Participants who scored low on the core screening questions (see "Assessment Design" below) were given ALSA instead of the main assessment.

The Main Assessment. NAAL main assessment reports a separate score for each of three literacy

areas: prose literacy, document literacy, and quantitative literacy.

Prose literacy refers to the knowledge and skills needed to perform prose tasks—that is, to search, comprehend, and use continuous texts. Prose examples include editorials, news stories, brochures, and instructional materials.

Document literacy refers to knowledge and skills needed to perform document tasks—that is, to search, comprehend, and use continuous texts. Document examples include job applications, payroll forms, transportation schedules, maps, tables, and drug or food labels.

Quantitative literacy refers to the knowledge and skills needed to perform prose tasks—that are, to identify and perform computations, either alone or sequentially, using numbers embedded in printed materials. Examples include balancing a check book, computing a tip, completing and order form, or determining the amount of interest on a loan from an advertisement.

Periodicity

The 2003 NAAL results are comparable to those of the 1992 NALS, and for young adults 21 to 25 years old, to the 1985 young adult literacy assessment.

2. USES OF DATA

NAAL data provide vital information to policymakers, business and labor leaders, researchers, and citizens. The survey results can be used to

- describe the levels of literacy demonstrated by the adult population as a whole and by adults in various subgroups (e.g., those targeted as at risk, prison inmates, and older adults);
- characterize adults' literacy skills in terms of demographic and background information (e.g., reading characteristics, education, and employment experiences);
- profile the literacy skills of the nation's workforce;
- compare assessment results from the current study with those from the 1992 NALS;
- interpret the findings in light of information-processing skills and strategies, so as to inform curriculum decisions concerning adult education and training; and
- increase our understanding of the skills and knowledge associated with living in a technological society.

3. KEY CONCEPTS

NAAL is designed to measure functional English literacy. The assessment measures how adults use printed and written information to adequately function at home, in the workplace, and in the community.

Since adults use different kinds of printed and written materials in their daily lives, NAAL measures three types of literacy—prose, document, and quantitative—and reports a separate scale score for each of these three areas. By measuring literacy along three scales, instead of just one, NAAL can provide more comprehensive data on literacy tasks and literacy skills associated with the broad range of printed and written materials adults use.

Prose Literacy

The prose literacy scale measures the knowledge and skills needed to perform prose tasks (i.e., to search, comprehend, and use continuous texts). Examples include editorials, news stories, brochures, and instructional materials.

Document Literacy

The document literacy scale measures the knowledge and skills needed to perform document tasks (i.e., to search, comprehend, and use non-continuous texts in various formats). Examples include job applications, payroll forms, transportation schedules, maps, tables, and drug or food labels.

Quantitative Literacy

The quantitative literacy scale measures the knowledge and skills required to perform quantitative tasks (i.e., to identify and perform computations, either alone or sequentially, using numbers embedded in printed materials). Examples include balancing a checkbook, figuring out a tip, completing an order form, or determining the amount of interest on a loan from an advertisement.

In addition to the prose, document, and quantitative literacy scales, the 2003 assessment included a health literacy scale. The health literacy scale contains prose, document, and quantitative items with health-related

content. The items fall into three areas: clinical, prevention, and navigation of the health system.

4. SURVEY DESIGN

Data collection for the main NAAL study and the concurrent state assessment, SAAL, was conducted in 2003 using in-person household interviews. Over 18,000 adults participated, selected from a sample of over 35,000 households that represented the entire U.S. household population age 16 and over—about 222 million Americans (U.S. Census Bureau, Current Population Survey 2003). In addition, approximately 1,200 inmates from 110 federal and state prisons were assessed in early 2004 for the prison component, which provides separate estimates of literacy levels for the incarcerated population.

All household participants received an incentive payment of $30 in an effort to increase both the representativeness of the sample and the response rate. Black and Hispanic households were oversampled at the national level to ensure reliable estimates of their literacy proficiencies. Special accommodations were made for adults with disabilities or with limited English proficiency.

Target Population

The target population for the national household sample consisted of adults 16 and older in the 50 states and the District of Columbia who, at the time of the survey, resided in private households or college dormitories. The target population for the supplemental state household sample consisted of individuals 16 to 64 years of age who, at the time of the survey, resided in private households or college dormitories in the participating state. The target population for the prison sample consisted of inmates age 16 and older in state and federal prisons at the time of the survey; those held in local jails, community-based facilities, or other types of institutions were not included.

Sample Design

The 2003 NAAL included two samples: (1) adults age 16 and older living in households (99 percent of the entire NAAL sample, weighted); and (2) inmates age 16 and older in state and federal prisons (1 percent of the entire NAAL sample, weighted). Each sample was weighted to represent its share of the total population of the United States, and the samples were combined for reporting.

Household sample. The 2003 NAAL household sample included a nationally representative probability sample of 35,000 households. The household sample was selected on the basis of a four-stage, stratified area sample: (1) primary sampling units (PSUs) consisting of counties or groups of contiguous counties; (2) secondary sampling units (referred to as segments) consisting of area blocks; (3) housing units containing households; and (4) eligible persons within households. Person-level data were collected through a screener, a background questionnaire, the literacy assessment, and the oral module.

Six states—Kentucky, Maryland, Massachusetts, Missouri, New York, and Oklahoma—purchased additional cases in their states to allow reporting at the state level. A single area sample was selected for the national NAAL sample, and additional samples were selected for the six states participating in the SAAL. For each sample, the usual procedures for area sampling were followed: a stratified probability proportionate to size design was used for the first two stages, and systematic random samples were drawn in the last two stages.

A key feature of the national NAAL sample was the oversampling of Black and Hispanic adults, which was accomplished by oversampling segments with high concentrations of these groups. The SAAL samples did not include any oversampling of Black, Hispanic, or other racial/ethnic groups.

Although integrating the NAAL and SAAL samples at the design stage would have been more effective statistically, the states agreed to participate after the NAAL sample design and selection process had been finalized. Therefore, the approach used in the 1992 NALS was followed: selecting the SAAL samples independently of the NAAL sample and combining the samples at the estimation phase by using composite estimation.

Stage one sampling. The first stage of sampling was the selection of PSUs, which consisted of counties or groups of counties. PSUs were formed within state boundaries, which gave an improved sample for state-level estimation. One PSU was selected per stratum by using probabilities proportionate to their population within households, except in Maryland and Massachusetts, where samples of segments were selected as the first-stage units. One hundred PSUs were selected for the national sample, and 54 PSUs were selected in Kentucky, Missouri, New York, and Oklahoma. Maryland and Massachusetts had too few PSUs from which to sample; therefore, segments were selected in the first stage of sampling. After selecting the segments, 20 area clusters (quasi-PSUs) were

created for Maryland and Massachusetts by grouping the selected segments into 20 geographically clustered areas to facilitate a cost-efficient approach to data collection. The true first-stage sample size is much larger because a total of 323 first-stage units (i.e., segments) were selected in Maryland and Massachusetts. Fourteen PSUs were selected for both the national NAAL and the SAAL samples; hence, the sample included a combined total of 160 unique PSUs.

Stage two sampling. In the second stage of sampling, segments (census blocks or groups of blocks) within the PSUs were selected with a probability proportionate to size; the measure of size for a segment was a function of the number of year-round housing units within the segment. In the national sample, the Black and Hispanic populations were sampled at a higher rate than the remainder of the population to increase their sample size, whereas the state samples used no oversampling. Oversampling in the national sample was accomplished by oversampling the segments in which Black and Hispanic adults accounted for 25 percent or more of the population. There were 2,000 segments selected for the national sample and 861 segments selected across the SAAL samples, with a total of 2,800 unique segments selected across the national and six SAAL samples. (Two segments were selected for both the NAAL and SAAL samples.)

Stage three sampling. In the third stage of sampling, housing units were selected with equal probability within each segment, except for White households within high percentage of Black, Hispanic, and other race/ethnicity segments in the national component. These national sample households were subsampled after screening so that the sampling rates for White persons would be about the same in the high percentage of Black, Hispanic, and other race/ethnicity segments as in other segments. The overall sample size of housing units took into account expected losses owing to vacant housing units, units that were not housing units, and expected response rates.

Stage four sampling. The fourth stage of selection involved listing the age-eligible household members (age 16 and older) for each selected household. Subsequently, one person was selected at random within households with three or fewer eligible persons, and two persons were selected if the household had four or more eligible persons. The listing and selection of persons within households were performed with the CAPI system.

Of the 35,000 sampled households, 4,700 were either vacant or not a dwelling unit, resulting in a sample of 31,000 households. A total of 25,000 households completed the screener, which was used to select survey respondents. The final screener response rate was 81 percent (weighted).

On the basis of the screener data, 24,000 respondents age 16 and older were selected to complete the background questionnaire and the assessment; 18,000 actually completed the background questionnaire. Of the 5,500 respondents who did not complete the background questionnaire, 360 were unable to do so because of a literacy-related barrier, either the inability to communicate in English or Spanish (the two languages in which the background questionnaire was administered) or a mental disability.

The final response rate for the background questionnaire—which included respondents who completed the background questionnaire and respondents who were unable to complete the background questionnaire because of language problems or a mental disability—was 77 percent (weighted). Of the 18,000 adults age 16 and older who completed the background questionnaire, 17,000 completed at least one question on each of the three scales—prose, document, and quantitative—measured in the adult literacy assessment. An additional 149 were unable to answer at least one question on each of the three scales for literacy-related reasons or a mental disability. The final response rate for the literacy assessment—which included respondents who answered at least one question on each scale plus the 150 respondents who were unable to do so because of language problems or a mental disability—was 97 percent (weighted).

Cases were considered complete if the respondent completed the background questionnaire and at least one question on each of the three scales or if the respondent was unable to answer any questions because of language issues (an inability to communicate in English or Spanish) or a mental disability. All other cases that did not include a complete screener, a background questionnaire, and responses to at least one question on each of the three literacy scales were considered incomplete or missing. Before imputation, the overall response rate for the household sample was 60 percent (weighted).

Imputation for nonresponse. For respondents who did not complete any literacy tasks on any scale, no information is available about their performance. Completely omitting these individuals from the analyses would have resulted in unknown biases in

estimates of the literacy skills of the national population because refusals cannot be assumed to have occurred randomly. For 860 respondents who answered the background questionnaire but refused to complete the assessment for reasons other than language issues or a mental disability, regression-based imputation procedures were applied to impute responses to one assessment item on each scale by using the NAAL background data on age, gender, race/ethnicity, education level, country of birth, census region, and metropolitan statistical area status.

On the prose and quantitative scales, a response was imputed for the easiest task on each scale. On the document scale, a response was imputed for the second easiest task because that task was also included on the health literacy scale. In each of the logistic regression models, the estimated regression coefficients were used to predict missing values of the item to be imputed. For each nonrespondent, the probability of answering the item correctly was computed and then compared with a randomly generated number between 0 and 1. If the probability of getting a correct answer was greater than the random number, the imputed value for the item was 1 (correct); otherwise, it was 0 (wrong). In addition, a wrong response on each scale was imputed for 65 respondents who started to answer the assessment, but were unable to answer at least one question on each scale because of language issues or a mental disability.

The final household reporting sample—including the imputed cases—consisted of 18,000 respondents. These 18,000 respondents include the 17,000 respondents who completed the background questionnaire and the assessment; the 860 respondents who completed the background questionnaire, but refused to do the assessment for non-literacy-related reasons (and have imputed responses to one item on each scale); and the 70 respondents who started to answer the assessment items, but were unable to answer at least one question on each scale because of language issues or a mental disability. After including the cases for which responses to the assessment questions were imputed, the weighted response rate for the household sample was 62 percent (18,000 cases with complete or imputed data and an additional 440 cases that had no assessment data because of language issues or a mental disability).

Prison sample. The 2003 assessment also included a nationally representative probability sample of inmates in state and federal prisons. The target population for the prison sample consisted of inmates age 16 and older from state and federal prisons in the United States. The sampling frame was created primarily from two data sources: the Bureau of Justice Statistics 2000 Census of State and Federal Adult Correctional Facilities (referred to in the following text as the Prison Census) and the 2003 Directory of Correctional Facilities of the American Correctional Association (ACA).

The facility universe for the NAAL Prison Component was consistent with the Prison Census. As defined for the Prison Census, the 2003 NAAL target population included the following types of state and federal adult correctional facilities: prisons; prison farms; reception, diagnostic, and classification centers; road camps; forestry and conservation camps; youthful offender facilities (except in California); vocational training facilities; drug and alcohol treatment facilities; and state-operated local detention facilities in Alaska, Connecticut, Delaware, Hawaii, Rhode Island, and Vermont. Facilities were included in the NAAL Prison Component if they were:

- staffed with federal, state, local, or private employees;
- designed to house primarily state or federal prisoners;
- physically, functionally, and administratively separate from other facilities; and
- in operation between September 2003 and March 2004.

Specifically excluded from the NAAL Prison Component were:

- privately operated facilities that were not exclusively for state or federal inmates;
- military facilities;
- Immigration and Naturalization Service facilities;
- Bureau of Indian Affairs facilities;
- facilities operated and administered by local governments, including those housing state prisoners;
- facilities operated by the U.S. Marshals Service, including the Office of the Detention Trustee;

- hospital wings and wards reserved for state prisoners; and

- facilities housing only juvenile offenders.

Even though they contain inmates up to age 21, juvenile facilities were excluded from NAAL for two reasons: (1) to remain consistent with the facilities listed in the Prison Census; and (2) to promote cost efficiency because it would not have been cost-effective to visit these facilities to sample the small number of inmates 16 years of age and older.

Inmate sampling frames were created by interviewers at the time they visited the prisons. The frame consisted of all inmates occupying a bed the night before inmate sampling was conducted.

Approximately 110 prisons were selected to participate in the adult literacy assessment. The final prison response rate was 97 percent (weighted). Among the inmates in these prisons, 1,300 inmates ages 16 and older were randomly selected to complete the background questionnaire and assessment. Of these 1,300 selected inmates, 1,200 completed the background questionnaire. Of the 140 inmates who did not complete the background questionnaire, about 10 were unable to do so because of a literacy-related barrier (either the inability to communicate in English or Spanish) or a mental disability.

The final response rate for the prison background questionnaire—which included respondents who completed the background questionnaire and respondents who were unable to complete the background questionnaire because of language problems or a mental disability—was 91 percent (weighted). Of the 1,200 inmates who completed the background questionnaire, 1,100 completed at least one question on each of the three scales—prose, document, and quantitative—measured in the adult literacy assessment. An additional 10 inmates were unable to answer at least one question on each of the three scales for literacy-related reasons. The final response rate for the literacy assessment—which included respondents who answered at least one question on each scale or were unable to do so because of language problems or a mental disability—was 99 percent (weighted).

The same definition of a complete case used for the household sample was also used for the prison sample, and the same rules were followed for imputation. Before imputation, the final response rate for the prison sample was 87 percent (weighted).

Imputation for nonresponse. One response on each scale was imputed on the basis of background characteristics for 30 inmates who completed the background questionnaire, but had incomplete or missing assessments for reasons that were not literacy related. The statistical imputation procedures were the same as for the household sample. The background characteristics used for the missing data imputation for the prison sample were prison security level, region of country/type of prison, age, gender, educational attainment, country of birth, race/ethnicity, and marital status. A wrong response on each scale was imputed for the inmates who started to answer the assessment, but were unable to answer at least one question on each scale because of language issues or a mental disability. The final prison reporting sample—including the imputed cases—consisted of 1,200 respondents. After the cases for which responses to the assessment questions were imputed were included, the weighted response rate for the prison sample was 88 percent (1,200 cases with complete or imputed data and an additional 20 cases that had no assessment data because of language issues or a mental disability).

Assessment Design

The NAAL interview was conducted in the order described below.

First, every respondent completed a background questionnaire that collected data on demographic, socioeconomic, and other factors associated with literacy.

Next, every respondent completed seven core screening questions, which were among the easiest in the assessment.

Similar in structure to the main NAAL assessment questions, the core questions determined whether a respondent's skills were sufficient to participate in the main NAAL assessment or if the individual should be routed to ALSA. Interviewers used a scoring rubric to code respondents' answers to each code question (e.g., "1" for correct, "2" for wrong, and "3" for no response). Interviewers entered the codes into a CAPI System, which selected respondents for ALSA using an empirically derived algorithm that predicts very low performance on the main NAAL. ALSA assessed the ability of the least literate adults to identify letters and numbers and to comprehend simple prose materials. Those participants who scored low on the basic core screening questions took ALSA instead of the main NAAL.

After completing either the main NAAL assessment booklet or ALSA, every respondent took FAN. FAN used speech-recognition software to assess adults' ability to decode and recognize words and to read with fluency.

Data Collection and Processing

Reference dates. Household data collection was conducted from March 2003 through February 2004; prison data collection was conducted from March through July 2004.

Data collection. Household interviews took place in respondents' homes; prison interviews generally took place in a classroom or library in the prison. Whenever possible, interviewers administered the background questionnaire and assessment in a private setting. Unless there were security concerns, a guard was not present in the room when inmates were interviewed.

Interviewers used a CAPI system programmed into laptop computers. The interviewers read the background questions from the computer screen and entered all responses directly into the computer. Skip patterns and follow-up probes for contradictory or out-of-range responses were programmed into the computer.

After completing the background questionnaire, respondents were handed a booklet with the assessment questions. The interviewers followed a script that introduced the assessment booklet and guided the respondent through the assessment.

Each assessment booklet began with the same seven screening questions. After the respondent completed the screening questions, the interviewer asked the respondent for the book and used an algorithm to determine, on the basis of the responses to the questions, whether the respondent should continue in the main assessment or be placed in ALSA. Three percent (weighted) and 5 percent (unweighted) of adults were placed in the ALSA.

ALSA is a performance-based assessment that allowed adults with marginal literacy to demonstrate what they could and could not do when asked to make sense of various forms of print. The ALSA started with simple identification tasks and sight words and moved to connected text, using authentic, highly contextualized material commonly found at home or in the community.

Respondents were routed to an alternative assessment (ALSA) based on their performance on the seven easy screening tasks at the beginning of the literacy assessment. Because the ALSA respondents answered most, or all, of these questions incorrectly, if they were place on the NAAL scale, they would have been classified on the NAAL scale as below basic level on the health scale.

A respondent who continued in the main assessment was given back the assessment booklet, and the interviewer asked the respondent to complete the tasks in the booklet and guided the respondent through them. The main assessment consisted of 12 blocks of tasks with approximately 11 questions in each block, but each assessment booklet included only 3 blocks of questions. The blocks were spiraled so that across the 26 different configurations of the assessment booklet, each block was paired with every other block and each block appeared in each of the three positions (first, middle, last) in a booklet.

For ALSA interviews, the interviewer read the ALSA script from a printed booklet and classified the respondent's answers into the response categories in the printed booklet. ALSA respondents were handed the materials they were asked to read.

Following the main assessment or ALSA, all respondents were administered FAN (the oral fluency assessment). Respondents were handed a booklet with passages, number lists, letter lists, word lists, and pseudoword lists to read orally. Respondents read into a microphone that recorded their responses on the laptop computer.

Accommodations. With the passage of the Americans with Disabilities Act and the growth of America's immigrant population, assessment programs like NAAL must consider issues of inclusion and accommodation. The 2003 NAAL provided for two types of accommodations—administrative and language.

Administrative accommodations were made for adults with disabilities. First, NAAL is inherently accommodating because the assessment was conducted one-on-one in the respondent's home. Second, all respondents with disabilities received additional time to complete the assessment, if necessary.

Language accommodations were made for adults with limited English proficiency or whose primary language is not English. Questions on the background questionnaire were available in either English or Spanish. In addition, instructions for FAN, ALSA, and the core screening test questions were given in either English or Spanish. However, the stimulus materials

for these questions were in English since NAAL's main objective is to assess literacy in English.

Results are reported separately for non-native speakers of English and compared to the results of native speakers of English. Thus, the unique needs of English as a Second Language (ESL) adults may be better understood by researchers, policymakers, and practitioners.

Data processing. The NAAL assessment questions were open-ended and thus required scoring by trained scorers. NAAL experts have developed scoring rubrics that detail the rules necessary for scoring each assessment question.

In order to make NAAL scores meaningful, the scores were grouped into performance levels to provide information that could more easily be understood and used by the public and policymakers. The performance levels were developed to characterize the status of English language literacy of American adults and include the following: nonliterate in English, below basic, basic, intermediate, and proficient literacy. For reporting purposes adults classified as nonliterate in English are included in the below basic literacy level. The 2003 NAAL performance levels are different from the five levels NCES used to report NALS results in 1992. However, in order to make comparisons across years, the 1992 data were reanalyzed and the new performance levels were applied to the 1992 data.

NAAL scoring is designed to measure adults' abilities to perform literacy tasks in everyday life. Since adults are likely to make mistakes as they interact with printed and written material, NAAL scorers make allowances for partial responses and writing errors.

While most responses are either correct or incorrect, a response can be partially correct if the information provided is still useful in accomplishing the task. For example, a respondent who writes the wrong product price on a catalog order form could receive partial credit, because in real life such a minor error would not necessarily result in the placement of an incorrect order (since other information is provided, such as product name and price). However, if a respondent miswrites a social security number on a government application form, such an error would not receive partial scoring.

Similarly, responses containing writing errors—grammatical and spelling errors, use of synonyms, incomplete sentences, or circling instead of writing the correct answer—are scored as correct as long as the overall meaning is correct and the information provided accomplishes the task. However, if a respondent is filling out a form and writes the answer on the wrong line, or if, for a quantitative task, the calculation is right but the respondent writes the wrong answer in the blank, then the response is scored as incorrect.

During the task development stage, scoring experts developed scoring rubrics that detailed the rules for scoring each assessment question. To ensure that all assessment questions were scored accurately, NAAL scoring rubrics underwent several stages of verification both before and after the assessment was administered.

Before the main NAAL study began, a field test of about 1,400 adults was conducted to help identify and screen out problems with the scoring rubrics, such as alternative correct responses and scoring rubrics that are difficult to implement consistently (thus leading to low rates of interrater reliability).

After the main study ended, a sample of responses from the household and prison interviews was scored using the scoring rubrics. As the test developers scored the sample responses, they made adjustments to the scoring rubrics to reflect the kinds of responses adults gave during the assessment. Together, these sample responses and the revised scoring rubrics were used in training the scorers who scored the entire assessment.

In a group setting, scorers were trained to recognize each task and its corresponding scoring rubric, as well as sample responses that are representative of correct, partially correct, and incorrect answers. After group training, readers scored numerous practice questions before they began to score actual booklets.

To ensure that readers were scoring accurately, 50 percent of the assessment questions were subject to a second interrater reliability check, in which a second reader scored the booklet and the scores of the first and second readers were compared. Interrater reliability is the percentage of times two readers agree exactly in their scores. (In 1992, the average percentage of agreement was 97 percent.) Any batch of questions that exceeded a low level of scoring mistakes was sent back to the scorers for corrections. Also, the scoring supervisor discussed the discrepancy with the scorers involved. Quality control procedures like this ensured reliability of the scoring.

Performance levels. Performance levels are important because they provide the ability to group people with similar literacy scores into a relatively small number of categories of importance to the adult education community, much like grouping students with similar scores on a test into various letter grades (e.g., A or B).

A benefit of having performance levels is that they enable NAAL to characterize American adults' relative literacy strengths and weaknesses by describing the nature and difficulty of the literacy tasks that participants at each level can perform with a reasonably high rate of success.

Performance levels were determined in response to a request from NCES to the National Research Council (NRC), which convened a Committee on Performance Levels for Adult Literacy. The committee's goal was to do the following in an open and public way: evaluate the literacy levels used by NAAL's 1992 predecessor survey, and recommend a set of performance levels that could be used in reporting the 2003 results and also be applied to the 1992 results in order to make comparisons across years.

New performance levels. After reviewing information about the 1992 and 2003 assessments as well as feedback from stakeholders (e.g., adult literacy practitioners), the NRC committee specified a new set of performance levels intended to correspond to four policy-relevant categories of adults, including adults in need of basic adult literacy services. The next step was to determine the score ranges to be included in each level for each of the three NAAL literacy scales—prose, document, and quantitative literacy.

Score ranges. To determine the score ranges for each level, the committee decided to use the "bookmark" method. Initial implementation of the method involved describing the literacy skills of adults in the four policy-relevant levels, and holding two sessions with separate panels of "judges" consisting of adult literacy practitioners, officials with state offices of adult education, and others. One group of judges focused on the 1992 assessment tasks and the other group focused on the 2003 assessment tasks.

Bookmarks. For each literacy area (prose, document, and quantitative), the judges were given, in addition to descriptions of the performance levels, a booklet of assessment tasks arranged from easiest to hardest. The judges' job was to place "bookmarks" in the set of tasks that adults at each level were "likely" to get right. The term "likely" was defined as "67 percent of the time," or two out of three times, and statistical procedures were used to determine the score associated with a 67 percent probability of performing the task correctly. The bookmarks designated by the judges at the two sessions were combined to produce a single bookmark-based cut score for each performance level on each of the three literacy scales.

Quasi-contrasting groups approach. To refine the bookmark-based cut scores, which indicated the lowest score to be included in each performance level, the committee used a procedure it termed the "quasi-contrasting groups approach." The committee compared the 2003 bookmark-based cut scores with the 1992 scores associated with various background variables, such as educational attainment. The criterion for selecting the background variables was potential usefulness for distinguishing between adjacent performance levels, such as basic and below basic (e.g., having some high school education vs. none at all; reporting that one reads well vs. not well; reading a newspaper sometimes vs. never reading a newspaper; reading at work sometimes or more often vs. never reading at work).

In each case, the midpoint between the average scores of the two adjacent performance levels (below basic and basic; basic and intermediate; intermediate and proficient) was calculated and averaged across the variables that provided contrast between the groups. The committee developed a set of rules and procedures for deciding when and how to make adjustments to the bookmark cut scores when the cut scores associated with the selected background variables were different from the bookmark-based scores.

Nonliterate in English classification. The NRC committee recommended that NCES distinguish a fifth group of adults with special importance to literacy policy—those who are nonliterate in English. As originally defined by the committee, this category consisted of adults who performed poorly on a set of easy screening tasks in 2003 and therefore were routed to an alternative assessment for the least literate adults (i.e., ALSA). Because the 1992 assessment included neither the alternative assessment nor the 2003 screening tasks, adults in this category cannot be identified for 1992.

To provide a more complete representation of the adult population that is nonliterate in English, NCES expanded the category to include not only the 3 percent of adults who took the alternative assessment, but also the 2 percent who were unable to be tested at all because they knew neither English nor Spanish (the other language spoken by interviewers). Thus, as defined by NCES, the category included about 5 percent of adults in 2003.

Refinements made before using the new levels. The new performance levels were presented to NCES as recommendations. Having accepted the general recommendations, NCES incorporated a few refinements before using the levels to report results. First, NCES changed the label of the top category from advanced to proficient because the term "proficient"

better conveys how well the upper category of adults performs. Second, NCES added sample tasks from the 2003 assessment to illustrate the full range of tasks that adults at each level can perform, as well as a brief (one-sentence) summary description for each level to enhance public understanding. Third, as outlined in the previous paragraph, NCES included additional adults in the "nonliterate in English" category.

Estimation Methods

Weighting. As discussed above, NAAL included both a household sample and a prison sample. The household sample was further divided into the cases selected for the national sample and the additional cases selected in the six SAAL states. Weighting was done separately for the household and prison samples. However, the weights were developed so that the two samples could be used together in a combined sample.

Household sample weighting. Differential probabilities of selection into the NAAL household sample were adjusted by computing base weights for all adults selected into the sample. The base weight was calculated as the reciprocal of a respondent's final probability of selection. The weights were adjusted for nonresponse at both the screener level and the background questionnaire level. Additionally, trimming procedures were followed to reduce the impact of extreme weights. The background questionnaire weighting steps were done separately for the national and SAAL household samples, and each sample was calibrated separately to population estimates based on 2003 Current Population Survey (CPS) data. To combine the NAAL and SAAL household samples, composite weights were calculated for the respondents in the six participating SAAL states and the respondents in the national NAAL household sample in these six states. The composite weights were adjusted through poststratification and raking to match the 2003 CPS data.

Prison sample weighting. The prison component weighting consisted of four main steps. First, prison base weights were constructed using the probability of selection for each prison into the sample. Then, a nonresponse adjustment was made to the prison base weights to account for nonparticipating prisons. Next, inmate base weights were calculated using the prison nonresponse-adjusted weight and the within-prison sampling rate. Finally, the inmate base weights were raked to Bureau of Justice Statistics control totals to account for inmate nonresponse and noncoverage.

Variance estimation. A complex sample design was used to select assessment respondents. The properties of a sample selected through a complex design can be very different from those of a simple random sample. (In a simple random sample, every individual in the target population has an equal chance of selection and the observations from different sampled individuals can be considered to be statistically independent of one another.) Sampling weights should be used to account for the fact that the probabilities of selection were not identical for all respondents. All population and subpopulation characteristics based on the NAAL data should use sampling weights in their estimation.

Since the respondents were selected using complex sample design, conventional formulas for estimating sampling variability that assume simple random sampling (and, hence, independence of observations) are inappropriate. Standard errors calculated as though the data had been collected from a simple random sample would generally underestimate sampling errors. Therefore, the properties of the complex data collection design should be taken into account during the analysis of the data.

Scaling. Each respondent to NAAL received a booklet that included 3 of the 13 assessments blocks. Because each respondent did not answer all of the NAAL items, item response theory (IRT) methods were used to estimate average scores on the health, prose, document, and quantitative literacy scales; a simple average percent correct would not allow reporting results that were comparable for all respondents. IRT models calculate the probability of answering a question correctly as a mathematical function of proficiency or skill. The main purpose of IRT analysis is to provide a common scale on which performance on some latent trait can be compared across groups, such as those defined by sex, race/ethnicity, or place of birth.

IRT models assume that an examinee's performance on each item reflects characteristics of the item and characteristics of the examinee. All models assume that all items on a scale measure a common latent ability or proficiency dimension (e.g., prose literacy) and that the probability of a correct response on an item is uncorrelated with the probability of a correct response on another item, given fixed values of the latent trait. Items are measured in terms of their difficulty as well as their ability to discriminate among examinees of varying ability.

The assessment used two types of IRT models to estimate scale scores. The two-parameter logistic (2PL) model was used for dichotomous items (that is, items that are scored either right or wrong). For the partial credit items, the graded response logistic (GRL) model was used. The scale indeterminacy was solved by setting an origin and unit size to the

reported scale means and standard deviations from the 1992 assessment. Linear transformation was performed to transform the original scale metric to the final reporting metric.

IRT models predict the probability of success on an item for each point along the latent ability scale. By selecting a criterion value for this probability, a single scale point can be associated with the difficulty of each item, and visual displays can be constructed showing the difficulty of selected items along the scale. Such item maps aid in interpreting the assessment scales and in describing the performance levels. The assessment conformed to common industry practice by choosing the value of 0.67 as its response probability convention.

5. DATA QUALITY AND COMPARABILITY

The NAAL sampling design and weighting procedures assured that participants' responses could be generalized to the population of interest.

Sampling Error

In the 2003 survey, the use of a complex sample design, adjustments for nonresponse, and poststratification procedures resulted in dependence among the observations. Therefore, a jackknife replication method was used to estimate the sampling variance. The mean square error of replicate estimates around their corresponding full sample estimate provides an estimate of the sampling variance of the statistic of interest. The replication scheme was designed to produce stable estimates of standard errors for national and prison estimates as well as for the individual states.

The advantage of compositing the national and state samples during sample weighting was the increased sample size, which improved the precision of both the state and national estimates. However, biases could be present because the national PSU sample strata were not designed to maximize the efficiency of state-level estimates.

Nonsampling Error

The major source of nonsampling error in the 2003 NAAL was nonresponse error; special procedures were developed to minimize potential nonresponse bias based on how much of the survey the respondent completed. Other possible sources of nonsampling error were random measurement error and systematic error due to interviewers, coders, or scorers.

Coverage error. Coverage error could result from either the sampling frame of households or prisons being incomplete or from a household's or prison's failure to include all adults age 16 and older on the lists from which the sampled respondents were drawn. Special procedures and edits were built into NAAL to review both listers' and interviewers' ongoing work and to give any missed structures and/or dwelling units a chance of selection at data collection. However, just as all other household personal interview surveys have persistent undercoverage problems, the 2003 survey had problems in population coverage due to interviewers not gaining access to households in dangerous neighborhoods, locked residential apartment buildings, and gated communities.

Nonresponse error.

Unit nonresponse. Since three survey instruments—the screener, background questionnaire, and exercise booklet—were required for the administration of the survey, it was possible for a household or respondent to refuse to participate at the time of the administration of any one of these instruments. Because the screener and the background questionnaire were read to the survey participants in English or Spanish, but the exercise booklet required reading and writing in the English language, it was possible to complete the screener or background questionnaire but not the exercise booklet. Thus, response rates were calculated for each of the three instruments for the household samples. For the prison sample, there were only two points at which a respondent could not respond—at the administration of the background questionnaire or the exercise booklet.

For occupied households, "refusal or breakoff" was the most common explanation for nonresponse to the screener and the background questionnaire. The second most common explanation was "not at home after maximum number of calls." Nonresponse also resulted from language, physical, and mental problems. Housing units or individuals who refused to participate before any information was collected about them, or who did not answer a sufficient number of background questions, were not incorporated into the database. Because these individuals were unlikely to know that the survey intended to assess their literacy, it was assumed that their reason for not completing the survey was not related to their level of literacy.

There were reasons to believe that the literacy performance data were missing more often for adults with lower levels of literacy than for adults with higher levels. Field-test evidence and experience with surveys indicated that adults with lower levels of literacy were more likely than adults with higher levels either to decline to respond to the survey at all or to begin the

assessment but not complete it. Ignoring this pattern of missing data would have resulted in overestimating the literacy skills of adults in the United States. Therefore, to minimize bias in the proficiency estimates due to nonresponse to the literacy assessment, special procedures were developed to impute the literacy proficiencies of nonrespondents who completed fewer than five literacy tasks.

The household sample was subject to unit nonresponse from the screener, background questionnaire, literacy assessment, and oral module and to item nonresponse to background questionnaire items. Although all background questionnaire items had response rates of more than 85 percent, two stages of data collection—the screener and the background questionnaire—had unit response rates below 85 percent and thus required an analysis of the potential for nonresponse bias.

Table 13 presents a summary of the household response rate and table 14 presents a summary of the prison response rate.

Table 13. Weighted and unweighted unit response rates in the household sample of the National Assessment of Adult Literacy, by survey component: 2003

Component	Weighted response rate (percent)	Unweighted response rate (percent)
Screener	81.2	81.8
Background questionnaire	76.6	78.1
Literacy assessment	96.6	97.2
Overall response rate before imputation	60.1	62.1
Overall response rate after imputation	62.1	63.9

SOURCE: Greenberg, E., and Jin, Y. (2007). *2003 National Assessment of Adult Literacy: Public-Use Data File User's Guide* (NCES 2007-464). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Item nonresponse. For each background questionnaire, staff verified that certain questions providing critical information for weighting and data analyses had been answered, namely, education level, employment status, parents' level of education, race, and sex. If a response was missing, the case was returned to the field for data retrieval. Therefore, item response rates for completed background questionnaires were quite high, although they varied by type of question. Questions asking country of origin (first question in the booklet) and sex (last question in the booklet) had nearly 100 percent response rates, indicating that most respondents attempted to complete the entire questionnaire. Response rates were lower, however, for questions about income and educational background.

Table 14. Weighted and unweighted response rates in the prison sample of the National Assessment of Adult Literacy, by survey component: 2003

Component	Weighted response rate (percent)	Unweighted response rate (percent)
Prison	97.3	97.3
Background questionnaire	90.6	90.4
Literacy assessment	98.9	98.8
Overall response rate before imputation	87.2	86.8
Overall response rate after imputation	88.3	87.9

SOURCE: Greenberg, E., and Jin, Y. (2007). *2003 National Assessment of Adult Literacy: Public-Use Data File User's Guide* (NCES 2007-464). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

The CD-ROM: 2003 National Assessment of Adult Literacy Public-Use Data File User's Guide (Greenberg & Jin 2007) provides counts of item nonresponse. These, however, have to be considered in terms of the number of adults that were offered each task, because a great deal of the missing data is missing by design.

Nonresponse bias. NCES statistical standards require a nonresponse bias analysis when the unit response rate for a sample is less than 85 percent. The nonresponse bias analysis of the household sample revealed differences in the background characteristics of respondents who participated in the assessment compared with those who refused.

In bivariate unit-level analyses at the screener and background questionnaire stages, estimated percentages for respondents were compared with those for the total eligible sample to identify any potential bias owing to nonresponse. Although some statistically significant differences existed, the potential for bias was small because the absolute difference between estimated percentages was less than 2 percent for all domains considered. Multivariate analyses were conducted to further explore the potential for nonresponse bias by identifying the domains with the most differential response rates. These analyses revealed that the lowest response rates for the screener

were among dwelling units in segments with high median income, small average household size, and a large proportion of renters. The lowest response rates for the background questionnaire were among males age 30 and older in segments with high median income.

However, the variables used to define these areas and other pockets with low response rates were used in weighting adjustments. The analysis showed that weighting adjustments were highly effective in reducing the bias. The general conclusion was that the potential amount of nonresponse bias attributable to unit nonresponse at the screener and background questionnaire stages was likely to be negligible.

Measurement error. All background questions and literacy tasks underwent extensive review by subject area and measurement specialists, as well as scrutiny to eliminate any bias or lack of sensitivity to particular groups. Special care was taken to include materials and tasks that were relevant to adults of widely varying ages. During the test development stage, the tasks were submitted to test specialists for review, part of which involved checking the accuracy and completeness of the scoring guide. After preliminary versions of the assessment instruments were developed and after the field test was conducted, the literacy tasks were closely analyzed for bias or "differential item functioning." The goal was to identify any assessment tasks that were likely to underestimate the proficiencies of a particular subpopulation, whether it be older adults, females, or Black or Hispanic adults. Any assessment item that appeared to be biased against a subgroup was excluded from the final survey. The coding and scoring guides also underwent further revisions after the first responses were received from the main data collection.

Interviewer error checks. Several quality control procedures related to data collection were used during the field operation: an interviewer field edit, a complete edit of all documents by a trained field editor, validation of 10 percent of each interviewer's closeout work, and field observation of both supervisors and interviewers.

Coding/scoring error checks. In order to monitor the accuracy of coding, the questions dealing with country of birth, language, wages, and date of birth were checked in 10 percent of the questionnaires by a second coder. For the industry and occupation questions, 100 percent of the questionnaires were recoded by a second coder. Twenty percent of all the exercise booklets were subjected to a reader reliability check, which entailed a scoring by a second reader.

6. CONTACT INFORMATION

For content information about the NAAL project, contact:

Andrew Kolstad
Phone: (202) 502-7374
E-mail: andrew.kolstad@ed.gov

Mailing Address:
National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street, NW
Washington, DC 20006-5651

7. METHODOLOGY AND EVALUATION REPORTS

General

Greenberg, E., and Jin, Y. (2007). *2003 National Assessment of Adult Literacy: Public-Use Data File User's Guide* (NCES 2007-464). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y., and Dunleavy, E. (2007). *Literacy in Everyday Life: Results From the 2003 National Assessment of Adult Literacy* (NCES 2007-480). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Kutner, M., Greenberg, E., Jin, Y., and Paulsen, C. (2006). *The Health Literacy of America's Adults: Results From the 2003 National Assessment of Adult Literacy* (NCES 2006-483). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

White, S., and Dillow, S. (2006). *Key Concepts and Features of the 2003 National Assessment of Adult Literacy* (2006-471). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC

Chapter 21: Trends in International Mathematics and Science Study (TIMSS)

1. OVERVIEW

WORLDWIDE STUDY OF CLASSROOMS WITH AS MANY AS 58 COUNTRIES PARTICIPATING

TIMSS tests a variety of subject and content areas:

- Grade 4 math: Number, geometric shapes and measures, data display
- Grade 8 math: Number, algebra, geometry, data and chance
- Grade 4 science: Life, physical, and Earth science
- Grade 8 science: Earth science, biology, chemistry, and physics

The Trends in International Mathematics and Science Study (TIMSS) is a study of classrooms across the country and around the world. The National Center for Education Statistics (NCES), in the Institute of Education Sciences at the U.S. Department of Education, is responsible for the implementation of TIMSS in the United States. Beginning in 1995 and every 4 years thereafter, TIMSS has provided participating countries with an opportunity to measure students' progress in mathematics and science achievement. Studies of students, teachers, schools, curriculum, instruction, and policy issues are also carried out to understand the educational context in which learning takes place.

TIMSS represents the continuation of a long series of studies conducted by the IEA. The IEA conducted its First International Mathematics Study (FIMS) in 1964 and the Second International Mathematics Study (SIMS) in 1980–82. The First and Second International Science Studies (FISS and SISS) were carried out in 1970–71 and 1983–84, respectively. Since the subjects of mathematics and science are related in many respects and since there is broad interest among countries in students' abilities in both subjects, TIMSS began to be conducted as an integrated assessment of both mathematics and science.

In 1995, TIMSS collected data on grades 3 and 4 as well as grades 7 and 8, and the final grade of secondary school (grade 12 in the United States), with 42 countries participating. In 1999, data were collected only for 8th-grade students, with 38 countries participating. For TIMSS 2003 and 2007, data were collected on grades 4 and 8, with 46 countries participating in 2003 and 58 countries participating in 2007.

In addition to the math and science assessments given to students, supplementary information is obtained through the use of student, teacher, and school questionnaires. Also, in 1995 and 1999, further component studies were implemented, including benchmark and video studies.

The TIMSS 1999 Benchmarking Study included states and districts or consortia of districts from across the United States that chose to participate. These states and districts completed the assessments and questionnaires following the same procedures developed for the participating countries. They then used the findings to assess their comparative international standing and to evaluate their mathematics and science programs in an international context.

For the TIMSS Videotape Study, designed as the first study to collect videotaped records of classroom instruction, representative samples of 8th-grade mathematics classes in 1995 and 1999 and science classes in 1999 were drawn and one lesson in each of the participating classrooms was videotaped. The analysis provides a more detailed context for understanding mathematics and science teaching and learning in the classroom.

Purpose

TIMSS is designed to measure student performance in mathematics and science against what is expected to be taught in school. This focus on school curriculum allows for two broad questions to be addressed through TIMSS: (1) How do mathematics and science education environments differ across countries, how do student outcomes differ, and how are differences in these outcomes related to differences in mathematics and science education environments? (2) Are there patterns of relationships among contexts, inputs, and outcomes within countries that can lead to improvements in the theories and practices of mathematics and science education?

Components

TIMSS uses several types of instruments to collect data about students, teachers, schools, and national policies and practices that may contribute to student performance.

Written assessment. Assessments are developed to test students in various content areas within mathematics and science. For grade 4, the mathematics content areas are numbers; geometric shapes and measures; and data display. The grade 4 science content areas are Earth science; life science; and physical science. The grade 8 mathematics content areas are numbers; algebra; geometry; and data and chance. The grade 8 science content areas are biology; physics; chemistry; and Earth science.

In addition to being familiar with the mathematics and science content areas encountered in TIMSS, students are required to draw on a range of cognitive skills to successfully complete the assessment. TIMSS focuses on three cognitive domains in each subject: *knowing*, which covers the facts, procedures, and concepts students need to know; *applying*, which focuses on the ability of students to apply their knowledge and conceptual understanding to solve problems; and *reasoning*, which goes beyond solving routine problems to include unfamiliar situations and context that may require multi-step problem-solving.

After each TIMSS assessment cycle, approximately half of the items are publicly released, and replacement items that closely match the content of the original items are developed by international assessment and content experts. These new items are field tested and refined to the point where a variety of multiple choice and extended constructed-response items (i.e., items requiring written explanations from students) are chosen to be included in the TIMSS item pool.

Each student is asked to complete one booklet, made up of a subset of items taken from this item pool. No student answers all of the items in the item pool. The scoring of these booklets is accomplished through the use of a sophisticated and strict set of criteria that are implemented equally across all nations to ensure accuracy and comparability.

Student background questionnaire. Each student who takes the TIMSS assessment is asked to complete a questionnaire on issues including daily activities, family attributes, educational resources in the home, engagement in and beliefs about learning, instructional processes in the classroom, study habits, and homework.

Teacher questionnaire. The teacher questionnaire is given to the mathematics and science teachers of the students assessed in the study. These questionnaires ask about topics such as attitudes and beliefs about teaching and learning, teaching assignments, class size and organization, topics covered in class, the use of various teaching tools, instructional practices, professional preparation, and continuing development.

The teacher questionnaire is designed to provide information about the teachers of the students in the TIMSS student samples. The teachers who complete TIMSS questionnaires do not constitute a sample from any definable population of teachers. Rather, they represent the teachers of a national sample of students.

School questionnaire. The principal or head administrator is also asked to complete a questionnaire for the school focused on community attributes, personnel, teaching assignments, policy and budget responsibilities, curriculum, enrollment, student behavior issues, instructional organization, and mathematics and science courses offered.

Information collected from students, their teachers and schools is summarized in composite indices focused, in particular, on the relationship between mathematics and science achievement and the home, classroom, and school environment.

Curriculum questionnaire. The national research coordinator, or representative, of each participating country is asked to complete a questionnaire focused on the policies and practices supported at the national level that may contribute to student performance. In addition, because the mathematics and science topics covered in the assessment may not be included in all countries' curriculum, the national research coordinators are asked to indicate whether each topic

covered in TIMSS is included in their countries' intended curriculum through the fourth or eighth grade.

Encyclopedia. Beginning with TIMSS 2007, each participating country is asked to provide a written overview of the context in which mathematics and science instruction takes place, summarizing the structure of the education system, the mathematics and science curricula and instruction in primary and secondary grades, teacher education requirements, and the types of examinations and assessments employed to monitor success. The resulting chapters are compiled in *a publication entitled the TIMSS Encyclopedia*.

Videotape study. The 1995 TIMSS Videotape Study was designed as the first study to collect videotaped records of classroom instruction from national probability samples in Japan, Germany, and the United States in order to gather more in-depth information about the context in which learning takes place as well as to enhance understanding of the statistical indicators available from the main TIMSS study. An hour of regular classroom instruction was videotaped in a subsample of 8th-grade mathematics classrooms (except in Japan, where videotaping was usually done in a different class, selected by the principal) included in the assessment phase of TIMSS in each of the three countries.

The 1999 TIMSS Videotape Study was expanded in scope to examine national samples of 8th-grade mathematics and science instructional practices in seven nations: Australia, the Czech Republic, Hong Kong, Japan, the Netherlands, Switzerland, and the United States. Four countries—Australia, the Czech Republic, the Netherlands, and the United States—participated in both the mathematics and science components of the study. Hong Kong and Switzerland participated in only the mathematics component, and Japan in only the science component.

Curriculum studies. Continuing the approach of previous IEA studies, TIMSS addressed three conceptual levels of curriculum in 1995. The intended curriculum was composed of the mathematics and science instructional and learning goals as defined at the system level. The implemented curriculum was the mathematics and science curriculum as interpreted by teachers and made available to teachers. The attained curriculum was the mathematics and science content that students had learned and their attitudes toward these subjects. To aid in interpretation and comparison of results, TIMSS also collected extensive information about the social and cultural contexts for learning, many of which are related to variations among the education systems.

To gather information about the intended curriculum, mathematics and science specialists within each participating country worked section by section through curriculum guides, textbooks, and other curricular materials to categorize aspects of these materials in accordance with detailed specifications derived from TIMSS mathematics and science curriculum frameworks.

To collect data about how the curriculum was implemented in classrooms, TIMSS administered a broad array of questionnaires. These questionnaires were administered at the country level on decision making and organizational features within the education systems. The students who were tested answered questions pertaining to their attitudes toward mathematics and science, classroom activities, home background, and out-of-school activities. The mathematics and sciences teachers of sampled students responded to questions about teaching emphasis on the topics in the curriculum frameworks, instructional practices, textbook use, professional training and education, and their views on mathematics and science. The heads of schools responded to questions about school staffing and resources, mathematics and science course offerings, and support for teachers.

Ethnographic case studies. The case studies approach to understanding cultural differences in behavior has a long history in selected social science fields. Conducted only in 1995, the case studies were designed to focus on four key topics that challenge U.S. policymakers and to investigate how these topics were dealt with in the United States, Japan, and Germany: implementation of national standards; the working environment and training of teachers; methods for dealing with differences in ability; and the role of school in adolescents' lives. Each topic was studied through interviews with a broad spectrum of students, parents, teachers, and educational specialists. The ethnographic approach permitted researchers to explore the topics in a naturalistic manner and to pursue them in greater or lesser detail, depending on the course of the discussion. As such, these studies both validated and integrated the information gained from official sources with that obtained from teachers, students, and parents in order to ascertain the degree to which official policy reflected actual practice. The objective was to describe policies and practices in the nations under study that were similar to, different from, or nonexistent in the United States.

In three regions in each of the three countries, the research plan called for each of the four topics to be studied in the 4th, 8th, and 12th grades. The specific cities and schools were selected "purposively" to represent

different geographical regions, policy environments, and ethnic and socioeconomic backgrounds. Schools in the case studies were separated from schools in the main TIMSS sample. Where possible, a shortened form of the TIMSS test was administered to the students in the selected schools. The ethnographic researchers in each of the countries conducted interviews and obtained information through observations in schools and homes. Both native-born and nonnative researchers participated in the study to ensure a range of perspectives.

TIMSS benchmarking study. In 1999, 13 states and 14 districts or consortia of districts throughout the United States participated as their own "nations" in this project, following the same guidelines as the participating countries. The samples drawn for each of these states and districts were representative of the student population in each of these states and districts. The findings from this project allowed these jurisdictions to assess their comparative international standing and judge their mathematics and science programs in an international context.

NAEP/TIMSS linking study. A subsample of students who took the 2000 state National Assessment of Educational Progress (NAEP) mathematics and science assessment also took the 1999 TIMSS assessment. (See chapter 18 for more information on NAEP.) This provided an opportunity to compare students' performance on NAEP to their performance on TIMSS, and allowed for estimates of how states participating in the 2000 NAEP would have performed had they participated in TIMSS 1999. Results from the TIMSS 1999 Benchmarking Study were used to check the results of the linking study.

Periodicity

First conducted in 1995, TIMSS has been conducted every 4 years since then. Previous international math studies were conducted in 1964 and 1980–82; previous international science studies were conducted in 1970–71 and 1983–84.

2. USES OF DATA

The possibilities for specific research questions to be dealt with by TIMSS are numerous; however, the main research questions, focusing on the student, the school or classroom, and the national or international levels, are illustrated below:

- How much mathematics and science have students learned?
- How well are students able to apply mathematics and science knowledge to problem solving?
- What are students' attitudes toward mathematics and science?
- What do teachers teach in their classrooms?
- What methods and materials do teachers use in teaching mathematics and science, and how are they related to student outcomes?
- How strongly are students motivated to learn, in general, and to the learning of mathematics and science, in particular?
- What factors characterize the academic and professional preparation of teachers of mathematics and science?
- What are teachers' beliefs and opinions about the nature of mathematics and science (and about teaching them), and how are they related to the comparable opinions and attitudes of their students?
- What methods do teachers use to evaluate their students?
- If there are national curricula in a country, how specific are they, and what efforts are made to see that they are followed?

3. KEY CONCEPTS

Key terms related to TIMSS are described below.

National Desired Population. The stated objective in TIMSS is that the National Desired Population within each country be as close as possible to the International Desired Population, which is the target population. (See "Target Population" below under Section 4. Survey Design.) Using the International Desired Population as a basis, participating countries have to operationally define their populations for sampling purposes. Some national research coordinators have to restrict coverage at the country level, for example, by excluding remote regions or a segment of their country's education system. Thus, the National Desired Population sometimes differs from the International Desired Population.

4. SURVEY DESIGN

National Research Coordinators. This is an official from each participating country appointed to implement national data collection and processing in accordance with international standards. In addition to selecting the sample of students, national research coordinators are responsible for working with school coordinators, translating the test instruments, assembling and printing the test booklets, and packing and shipping the necessary materials to the sampled schools. They are also responsible for arranging the return of the testing materials from the school to the national center, preparing for and implementing the constructed-response item scoring, entering the results into data files, conducting on-site quality assurance observations for a 10 percent sample of schools, and preparing a report on survey activities.

Target Population

The International Desired Population for all countries is defined as follows:

- Grade 4: All students enrolled in the grade that represents 4 years of schooling, counting from the 1st year of the International Standard Classification of Education (ISCED) Level 1, providing that the mean age at the time of testing is at least 9.5 years. For most countries, the target grade should be the fourth grade or its national equivalent. All students enrolled in the target grade, regardless of their age, belong to the international desired target population.

- Grade 8: All students enrolled in the grade that represents 8 years of schooling, counting from the 1st year of ISCED Level 1, providing that the mean age at the time of testing is at least 13.5 years. For most countries, the target grade should be the eighth grade or its national equivalent. All students enrolled in the target grade, regardless of their age, belong to the international desired target population.

Thus, TIMSS uses a grade-based definition of the target population.

Sample Design

Each country participating in TIMSS, like the United States, is required to draw random samples of schools. In the United States, a national probability sample is drawn for each study that has resulted in over 500 schools and approximately 33,000 students participating in 1995, approximately 220 schools and 9,000 students participating in 1999, approximately 480 schools and almost 19,000 students in 2003, and approximately 500 schools and over 20,000 students in 2007. This sample design ensures the appropriate number of schools and students are participating to provide a representative sample of the students in a specific grade in the United States as a whole.

The TIMSS sample design for each country and population is intended to give a probability sample of all students within the target grades in the national school system (except for a small number of students allowed to be excluded as ineligible according to national criteria). Every eligible student in the country's school system has a chance of being selected, with a fixed probability of selection. These probabilities of selection are designed to be equal across eligible students as much as possible, but for a variety of reasons the probabilities of selection differ between students in most of the national samples.

Written assessment.

The TIMSS sample design is a two-stage stratified cluster sample, with schools as the first stage of selection and classrooms within schools as the second stage of selection. For the first time TIMSS 2007 included an optional third stage. The third-stage sampling units for TIMSS 2007 were students within sampled classrooms. Generally however, TIMSS chooses intact classrooms, so students are essentially chosen at the same stage as the classroom (i.e. the second stage).

Individual schools are selected with probability proportionate to size (PPS), size being the estimated number of students enrolled in the target grade. Prior to sampling, schools in the sampling frame can be assigned to a predetermined number of explicit or implicit strata. Substitution schools, selected to replace schools that refuse to participate, are identified simultaneously.

The classroom sampling design is intended to be an equal probability design with no subsampling in the classroom. However, a design based on a PPS sample of classrooms, with a fixed sample size of students selected within the sampled classroom, is permitted under the international guidelines. Exclusions can occur at the school level, the classroom level, or the student level. TIMSS participants are expected to keep such exclusions to no more than 10 percent of the National Desired Population.

The optional third-stage sampling unit for TIMSS 2007 was students within the sampled classrooms. While all students in a sampled classroom were to be selected for the assessment, it was possible for participating

countries to sample a subgroup of students after consultation with Statistics Canada, the organization serving as the sampling referee.

TIMSS standards for sampling precision require a minimum of 4,000 students to be assessed per grade. To meet the standard, at least 150 schools are selected per target population. However, the clustering effect of sampling classrooms rather than students is also considered in determining the overall sample size. Because the magnitude of the clustering effect is determined by the size of the cluster and the intraclass correlation, TIMSS produced sample-design tables showing the number of schools to sample for a range of intraclass correlations and minimum-cluster-size values. Some countries need to sample more than 150 schools. Countries, however, are asked to sample 150 schools even if the estimated number of schools necessary to be sampled is less than 150.

The schools in each explicit stratum (geographical region, public/private, etc.) are listed in order of the implicit stratification variables and then further sorted according to their measure of size. The stratification variables differ from country to country. Small schools are handled either through explicit stratification or through the use of pseudo-schools. In some very large countries, there is a preliminary sampling stage before schools are sampled in which the country is divided into primary sampling units.

In cases where a sampled school is unable to participate in the assessment, a replacement school is used. The replacement school is the next school on the ordered school-sampling list as the replacement for each particular sampled school. The school after that is a second replacement, should it be necessary. Using either explicit or implicit stratification variables and ordering of the school sampling frame by size ensures that any original sampled school's replacement has similar characteristics.

In the second stage of sampling, classrooms of students are sampled. Generally, in each school, one classroom is sampled from each target grade, although some countries opt to sample two classrooms at the upper grade in order to be able to conduct special analyses. Most countries test all students in selected classrooms, and in these instances the classrooms are selected with equal probabilities. A few participants use a design based on a PPS sample of classrooms, with a fixed sample size of students selected within the sampled classrooms. Participants with particularly large classrooms in their schools can decide to subsample a fixed number of students from each selected classroom. This is done using a simple random sampling method whereby all students in a sampled classroom are assigned equal selection probabilities.

In the United States, TIMSS 2007 used a two-stage stratified cluster sampling design based on the 2006 NAEP school sampling frame. The United States did not use the optional third stage of sampling (i.e. students within classrooms) for TIMSS 2007. (Time constraints related to recruitment activities required sample selection before the 2007 frame became available.) For this purpose the sampling frame, though not explicitly stratified, was implicitly stratified by four categorical variables: type of school (public or private); region of the country (Northeast, Central, West, Southeast); community type (eight levels); and percentage of Black, Hispanic, and other race/ethnicity students (above or below 15 percent of the student population).

The first stage of the design used a systematic PPS technique to select schools for the original sample. That is, schools were selected with a probability proportionate to the school's estimated enrollment of fourth- or eighth-grade students. Enrollment data for public schools were taken from the 2003–04 Common Core of Data (CCD), and data for private schools were taken from the 2003–04 Private School Universe Survey (PSS). For each original school selected, the two adjacent schools in the sampling frame, and within the same implicit stratum, were designated as the first and second replacement schools. The first substitute followed the original sample school in the frame listing and the second substitute preceded it. Substitute schools were designed to be used only if an original school refused to participate. In this situation the first substitute was to be contacted first, with the second substitute contacted only if the first substitute also refused to participate. Additionally, one sampled school was not allowed to substitute for another, and a given school could not be assigned to substitute for more than one sampled school.

An initial sample of 300 schools was selected at each grade level. Ineligible schools among these reduced the grade 4 sample to 290 schools and the grade 8 sample to 290 schools.

At each grade level, the U.S. sample design within schools consisted of an equal probability sample of two classrooms. In schools with a single eligible classroom, that classroom was selected with certainty. All eligible students in the classroom were designated to be in the sample (although generally the option for sub sampling did exist, there was no subsampling of students in the TIMSS 2007 U.S. sample).

Teacher questionnaire. The TIMSS database for each country includes questionnaire data from the teachers of the sampled classrooms, which can be linked to student assessment data in the classrooms. Any teacher linked as mathematics or science teacher to any assessed student is eligible to receive a questionnaire. The classroom sample is drawn from a listing of mathematics classrooms, so that in most situations only one mathematics teacher is linked to each sampled classroom. If this single teacher is also only linked to a single sampled classroom, then the teacher receives a questionnaire for that single classroom.

This straightforward one-to-one linking does not always hold, however. In some cases, teachers may teach both mathematics and science to students in a sampled classroom, making them eligible to receive questionnaires for both subjects.

For the U.S. TIMSS 2007 sample, a teacher was not asked to complete more than one questionnaire. In cases where a teacher taught both subject areas, the teacher was provided a specially designed questionnaire that included questions for both mathematics and science teachers.

In general, each country is allowed to develop its own methodology for this process of assigning subjects and classrooms to teachers when the links are not straightforward due to the presence of one to many (or many to one) mappings.

Assessment Design

TIMSS is a cooperative effort involving representatives from every country participating in the study. For TIMSS 2007, the development effort began with a revision of the frameworks that were used to guide the construction of the assessment. The frameworks were updated to reflect changes in the curriculum and instruction of participating countries. Extensive input from experts in mathematics and science education, assessment, and curriculum, and representatives from national education centers around the world contributed to the final shape of the frameworks used in 2007. Maintaining the ability to measure change over time is an important factor in constantly revising the frameworks.

Test development. As part of the TIMSS dissemination strategy, approximately one-half of the items at each grade are released for public use. To replace assessment items that have been released, countries submit items for review by subject-matter specialists, and additional items are written to ensure that the content, as explicated in the frameworks, is covered adequately. Items are reviewed by an international Science and Mathematics Item Review Committee and field tested in most of the participating countries. Results from the field tests are used to evaluate item difficulty, how well items discriminate between high- and low-performing students, the effectiveness of distracters in multiple-choice items, scoring suitability and reliability for constructed-response items, and evidence of bias toward or against individual countries or in favor of boys or girls.

Instrument design. TIMSS 2007 included booklets containing assessment items as well as questionnaires submitted to principals, teachers, and students. The assessment booklets were constructed such that not all of the students responded to all of the items, which is consistent with the design of other large-scale assessments, such as NAEP. To keep the testing burden to a minimum, and to ensure broad subject-matter coverage, TIMSS 2007 used a rotated block design that included both mathematics and science items. That is, students encountered both mathematics and science items during the assessment.

The U.S. 2007 fourth-grade assessment consisted of 14 booklets, each requiring approximately 72 minutes of response time. The 14 booklets were rotated among students, with each participating student completing only 1 booklet. The mathematics and science items were assembled into 14 blocks, or clusters, of items, with each block containing either mathematics or science items. The secure, or trend, items were included in 3 blocks, with the other 11 blocks containing replacement items. Each of the 14 booklets contained a total of 6 blocks.

The U.S. 2007 eighth-grade assessment consisted of 18 booklets, each requiring approximately 90 minutes of response time. The 18 booklets were rotated among students, with each participating student completing only 1 booklet. The mathematics and science items were assembled into 14 blocks, or clusters, of items, with each block containing either mathematics or science items. The secure, or trend, items were included in 3 blocks, with the other 11 blocks containing replacement items. Each of the 18 booklets contained a total of 4 blocks. As part of the design process, it was necessary to ensure that the booklets showed a distribution across the mathematics and science content domains as specified in the frameworks.

Data Collection and Processing

Data collection. TIMSS 2007 emphasized the use of standardized procedures in all countries. Each country collected its own data, based on comprehensive manuals and trainings provided by the international

project team to explain the survey's implementation, including precise instructions for the work of school coordinators and scripts for test administrators to use in testing sessions. Test administration in the United States was carried out by professional staff trained according to the international guidelines. School staff was asked only to assist with listings of students, identifying space for testing in the school, and specifying any parental consent procedures needed for sampled students.

Each country was responsible for conducting quality control procedures and describing this effort in the national research coordinator's report documenting procedures used in the study. In addition, the TIMSS International Study Center considered it essential to monitor compliance with the standardized procedures. National research coordinators were asked to nominate one or more persons unconnected with their national center, such as retired school teachers, to serve as quality control monitors for their countries. The International Study Center developed manuals for the monitors and briefed them in 2-day training sessions about TIMSS 2007, the responsibilities of the national centers in conducting the study, and their own roles and responsibilities.

Data entry and cleaning. Responsibility for data entry is taken by the national research coordinator from each participating country. The data collected for TIMSS 2007 were entered into data files with a common international format, as specified in the *Manual for Entering the TIMSS 2007 Data*. Data entry was facilitated by the use of common software available to all participating countries (WinDEM). The software facilitated the checking and correction of data by providing various data consistency checks. After data entry, the data were sent to the IEA Data Processing Center (DPC) in Hamburg, Germany, for cleaning. The DPC checked that the international data structure was followed; checked the identification system within and between files; corrected single-case problems manually; and applied standard cleaning procedures to questionnaire files. Results of the data cleaning process were documented by the DPC. This documentation was then shared with the national research coordinator with specific questions to be addressed. The national research coordinator then provided the DPC with revisions to coding or solutions for anomalies. The DPC then compiled background univariate statistics and preliminary classical and Rasch Item Analysis.

Estimation Methods

Once TIMSS data are scored and compiled, the responses are weighted according to the sample design and population structure and then adjusted for nonresponse. This ensures that countries' representation in TIMSS is accurately assessed. The analyses of TIMSS data for most subjects are conducted in two phases: scaling and estimation. During the scaling phase, Item Response Theory (IRT) procedures are used to estimate the measurement characteristics of each assessment question. During the estimation phase, the results of the scaling are used to produce estimates of student achievement (proficiency) in the various subject areas. The methodology of multiple imputations (plausible values) is then used to estimate characteristics of the proficiency distributions. Although imputation is conducted for the purpose of determining plausible values, no imputations are included in the TIMSS database.

Weighting. The TIMSS international design provides for two categories of sampling weights. The first category is designed to be used when schools, classrooms, or students are the unit of analysis. The second category is designed to be used in analyses where teachers, or both teachers and students, are the units of analysis.

First category. Sampling weights in the first category consist of school, classroom, and student weights, along with a combined student weight that is the product of these weights. The school weight is, essentially, the inverse of the probability of a school being sampled in the first stage of the sampling design. A school-level nonresponse adjustment is applied to compensate for any sampled schools that did not participate and were not replaced. This adjustment is calculated independently for each explicit stratum.

Classroom weights reflect the probability of the sampled classroom(s) being selected from among all the classrooms in the school at the target grade level. This classroom weight is calculated independently for each participating school. If a sampled classroom in a school does not participate, or if the participation rate among students in a classroom falls below 50 percent, a classroom-level participation adjustment is made to the classroom weight. If one (or more) selected classrooms in a school do not participate, the classroom participation adjustment is computed at the explicit stratum level rather than at the school level to reduce the risk of bias.

In the first category, student sampling weights are set at 1.0 since intact classrooms are sampled and each student in the sampled classrooms is certain of selection. A nonresponse adjustment is applied to adjust for sampled students who do not take part in the testing. This adjustment is calculated independently for each sampled classroom. An overall student sampling

weight is provided as well and is calculated as the product of the school, class, and student weights described above.

In addition, TIMSS provides "house" and "senate" weights, which are scaled versions of the overall student weight just described. The names are derived from an analogy with the U.S. legislative system. House weights are a set of weights based on the total sample size of each country, to be used when estimates across countries are computed or significance tests performed. The transformation of the weights will be different within each country, but in the end, the sum of the house-weight variables within each country will total to the sample size for that country. The house-weight variable is proportional to the total weight for that variable by the ratio of the sample size divided by the size of the population. These sampling weights can be used when the data user wants the actual sample size to be used in performing significance tests.

Senate weights are a set of weights based on a constant scalar, to be used when estimates across countries are computed or significance tests performed. The transformation of the weights will be different within each country, but in the end, the sum of the senate-weight variables within each country will total to a fixed value. The senate-weight variable, within each country, is proportional to the total weight for that variable by the ratio of the fixed value divided by the size of the population estimate. These sampling weights can be used when cross-national comparisons are required and the data user wants to have each country contribute the same amount to the comparison, regardless of the size of the population.

Second category. The teacher weight is a teacher-classroom weight and so is greater than 0 for a classroom only if the teacher filled out a questionnaire for that classroom. The teacher-classroom weight is equal to the sum of the student-teacher weights (see discussion below) for students linked to a classroom for a particular assessment.

Sampling weights in this second category are provided to facilitate analyses in which student and teacher data are analyzed together. TIMSS does not provide for a sample of teachers. Rather, the teachers in question are those who teach the sample of TIMSS students. As a consequence, analyses involving teachers have to be viewed as student-level analyses. Accordingly, teacher weights and student-teacher weights are derived from the overall student weight and are designed to accommodate the fact that students may have more than one teacher. Teacher weights are calculated by dividing the sampling weight for a student by the number of teachers that the student has. Separate mathematics and science student-teacher weights are developed by dividing the student sampling weight by, respectively, the number of mathematics teachers and the number of science teachers that the student has.

Scaling. TIMSS 1995, 1999, 2003, and 2007 used IRT procedures to produce scale scores that summarized the achievement results. With this method, the performance of a sample of students in a subject area or subarea can be summarized on a single scale or a series of scales, even when different students are administered different items. Because of the reporting requirements for TIMSS and because of the large number of background variables associated with the assessment, a large number of analyses have to be conducted. The procedures TIMSS uses for the analyses are developed to produce accurate results for groups of students while limiting the testing burden on individual students. Furthermore, these procedures provide data that can be readily used in secondary analyses. IRT scaling provides estimates of item parameters (e.g., difficulty, discrimination) that define the relationship between the item and the underlying variable measured by the test. IRT model parameters are estimated for each test question, with an overall scale being established as well as scales for each predefined content area specified in the assessment framework. For example, the TIMSS 2007 8^{th}-grade mathematics assessment had four scales describing mathematics content strands, and the science assessment had scales for four fields of science.

Imputation and plausible values. Although multiple imputation techniques are applied to create plausible values for student proficiency scores, with one exception, imputations were not generated for missing values in the TIMSS 2007 teacher, school, or student questionnaire data files. The single exception refers to a U.S.-only variable in the school file, the principal's report of the percentage of students eligible for free- or reduced-price lunch. For public schools, missing values for this variable were replaced by information obtained from the CCD. Analogous information was not available for private schools. Subsequently, analyses were undertaken to ensure that confidentiality was maintained.

During the scaling phase, plausible values are used to characterize scale scores for students participating in the assessment. To keep student burden to a minimum, TIMSS administers a limited number of assessment items to each student; too few to produce accurate content-related scale scores for each student. To account for this, for each student, TIMSS generates five possible content-related scale scores that represent

selections from the distribution of content-related scale scores of students with similar backgrounds who answer the assessment items the same way. The plausible-values technology is one way to ensure that the estimates of the average performance of student populations and the estimates of variability in these estimates are more accurate than those determined through traditional procedures, which estimate a single score for each student.

While constructing plausible values, careful quality control steps ensure that the subpopulation estimates based on these plausible values are accurate. Plausible values are constructed separately for each national sample. TIMSS uses the plausible-values methodology to represent what the true performance of an individual might have been, had it been observed. This is done by using a small number of random draws from an empirically derived distribution of score values based on the student's observed responses to assessment items and on background variables. Each random draw from the distribution is considered a representative value from the distribution of potential scale scores for all students in the sample who have similar characteristics and identical patterns of item responses. The draws from the distribution are different from one another to quantify the degree of precision (the width of the spread) in the underlying distribution of possible scale scores that could have caused the observed performance. The TIMSS plausible values function like point estimates of scale scores for many purposes, but they are unlike true point estimates in several respects. They differ from one another for any particular student, and the amount of difference quantifies the spread in the underlying distribution of possible scale scores for that student. Because of the plausible-values approach, secondary researchers can use the TIMSS data to carry out a wide range of analyses.

Scale anchoring. Beginning with TIMSS 2003, the percentage of students in each country performing at each of four international benchmarks of performance are reported. The benchmarks are selected to represent the range of performance of students internationally. The four benchmarks selected to represent points along the scale are advanced (set at 625), high (550), intermediate (475), and low (400). Using these points along the TIMSS scale, a scale anchoring analysis is conducted to describe student performance in terms of what they know and can do. The scale anchoring process involves a statistical component, which identifies assessment items that discriminate between points on the scale, and expert judgment, in which subject-matter specialists examine the items that anchor at different points along the scale and generalize about students' knowledge and understanding.

Future Plans

The next TIMSS data collection will take place in spring 2011. In addition, a new effort to link national and international assessments will be initiated in 2011 so that states can compare their own students' performance against international benchmarks. The linking study is intended to enable NCES to project state-level scores on the TIMSS using data from the National Assessment of Educational Progress (NAEP).

In the linking study, two representative national samples will be tested on their knowledge of mathematics and science by taking both the NAEP and TIMSS assessments. One sample of 10,000 eighth-graders will take combined test booklets in the winter of 2011 as part of NAEP. The other sample of 7,500 eighth-graders will take combined test booklets in the spring of 2011 as part of TIMSS. The relationships between the two assessments of mathematics and science that are found in these two samples will permit state-level projections of how the students in the 50 states and the District of Columbia that took NAEP would have performed in eighth-grade mathematics and science on TIMSS, with scores that can be compared to those of other countries. Data from a number of states that have agreed to administer TIMSS 2011 to state representative samples will be compared to the projected scores to ensure the accuracy of the linking projections.

5. DATA QUALITY AND COMPARABILITY

In addition to setting high standards for data quality, the TIMSS International Study Center has tried to ensure the overall quality of the study through a dual strategy of providing support to the national centers and performing quality control checks.

Despite the efforts taken to minimize error, any sample survey as complex as TIMSS has the possibility of error. Below is a discussion of possible sources of error in TIMSS.

Sampling Error

With complex sampling designs that involve more than the simple random sampling of students, as in the case of the stratified multistage design used in TIMSS 2007, where students were clustered within schools, there are several methods for estimating the sampling error of a

statistic that avoid the assumption of simple random sampling. One such method is the Jackknife Repeated Replication (JRR) technique. The particular application of the JRR technique used in TIMSS is termed a paired selection model because it assumes that the primary sampling units can be paired in a manner consistent with the sampling design, with each pair regarded as members of a pseudo-stratum for variance estimation purposes.

Following this first-stage sampling, there may be any number of subsequent stages of selection that may involve equal or unequal probability selection of the corresponding elements.

Imputation error. The variance introduced by imputation of missing data must be considered when using plausible values to estimate standard errors for proficiency estimates. The general procedure for estimating the imputation variance using plausible values is as follows: first estimate the statistic (*t*), each time using a different set of the plausible values (*M*). The statistics t_m can be anything estimable from the data, such as a mean, the difference between means, percentiles, etc. If all five plausible values in the TIMSS database are used, the parameter will be estimated five times, once using each set of plausible values. Each of these estimates will be called *t*, where *m=1, 2,…, 5*. Once the statistics are computed, the imputation variance is then computed as

$$Var_{imp} = (1 + 1/M)Var(t_m)$$

where *M* is the number of plausible values used in the calculation, and $Var(t_m)$ is the variance of the estimates computed using each plausible value.

Nonsampling Error

Due to the particular situations of individual TIMSS countries, sampling and coverage practices have to be adaptable, in order to ensure an internationally comparable population. As a result, nonsampling errors in TIMSS can be related both to coverage error and nonresponse. Measurement error is also a nontrivial issue in administering TIMSS, as different countries have different mathematics and science curricula. These potential sources of error are discussed in detail below.

Coverage error. The stated objective in TIMSS is that the effective population, the population actually sampled by TIMSS, be as close as possible to the International Desired Population. Yet, because a purpose of TIMSS is to study the effects of different international curricula and pedagogical methods on mathematics and science learning, participating countries have to operationally define their population for sampling purposes. Some national research coordinators have to restrict coverage at the country level, for example, by excluding remote regions or a segment of their country's education system. In these few situations, countries are permitted to define a National Desired Population that does not include part of the International Desired Population. Exclusions can be based on geographic areas or language groups.

Nonresponse error. Unit nonresponse error results from nonparticipation of schools and students. Weighted and unweighted response rates are computed for each participating country by grade, at the school level, and at the student level. Overall response rates (combined school and student response rates) are also computed.

The minimum acceptable school-level response rate for all countries, before the use of replacement schools, is set at 85 percent. This criterion is applied to the unweighted school-level response rate. However, both weighted and unweighted school-level response rates are calculated, with and without replacement schools. It is generally the case that weighted and unweighted response rates are similar.

Like the school-level response rate, the minimum acceptable student-level response rate is set at 85 percent for all countries. This criterion is applied to the unweighted student-level response rate. However, both weighted and unweighted student-level response rates are calculated. The weighted student-level response rate is the sum of the inverse of the selection probabilities for all participating students divided by the sum of the inverse of the selection probabilities for all eligible students.

Table 15 shows the unweighted unit level response rates for the data collections of 1995, 1999, 2003, and 2007 for grades 4 and 8.

Measurement error. Measurement error is introduced into a survey when its test instruments do not accurately measure the knowledge or aptitude they are intended to assess. The largest potential source of measurement error in TIMSS results from differences in the mathematics and science curricula across participating countries. In order to minimize the effects of measurement error, TIMSS carries out a special test called the *Test-Curriculum Matching Analysis*. Each country is asked to identify, for each item, whether the topic of the item is in the curriculum of the majority of the students.

Data Comparability

Through a careful process of review, analysis, and refinement, the assessment and questionnaire items are purposefully developed and field tested for similarity and for reliable comparisons between survey years. After careful review of all available data, including a test for reliability between old and new items, the TIMSS assessments are found to be very similar in format, content, and difficulty level across years.

Table 15. TIMSS unweighted unit-level response rates, by level, year, and grade: 1995, 1999, 2003, and 2007

Year and grade	School	Student	Overall
1995			
4th grade	86	94	81
8th grade	84	92	77
1999			
4th grade	†	†	†
8th grade	90	93	84
2003			
4th grade	83	95	78
8th grade	78	94	73
2007			
4th grade	89	95	84
8th grade	83	93	77

† Not available. TIMSS did not collect data from grade 4 in 1999.

SOURCE: Martin, M.O., and Kelly, D.L. (Eds.). (1998). *TIMSS Technical Report: Volume II: Implementation and Analysis, Primary and Middle School Years*. Boston College, International Study Center. Chestnut Hill, MA. Martin, M.O., Gregory, G.D., and Stemler, S.E. (Eds.). (2000). *TIMSS 1999 Technical Report*. Boston College, International Study Center. Chestnut Hill, MA. Martin, M.O., Mullis, I.V.S., and Chrostowski, S.J. (Eds.). (2004). *TIMSS 2003 Technical Report: Findings From IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades*. Boston College, International Study Center. Chestnut Hill, MA. Olson, J.F., Martin, M.O., and Mullis, I.V.S. (Eds.). (2008). *TIMSS 2007 Technical Report*. Boston College, International Study Center. Chestnut Hill, MA.

Findings from comparisons between the results of TIMSS, however, cannot be interpreted to indicate the success or failure of mathematics and science reform efforts within a particular country, such as the United States. International experts develop the TIMSS curriculum frameworks to portray the structure of the intended school mathematics and science curricula from many nations, not specifically the United States. Thus, when interpreting the findings, it is important to take into account the mathematics and science curricula likely encountered by U.S. students in school. TIMSS results are most useful, however, when they are considered in light of knowledge about education systems that include curricula, but also factors in trends in education reform, changes in school-age populations, and societal demands and expectations.

The ability to compare data across different countries constitutes a considerable part of the purpose behind TIMSS. As a result, it is crucial to ensure that items developed for use in one country are functionally identical to those used in other countries. Because questionnaires are originally developed in English and later translated into the language of each of the TIMSS countries, some differences do exist in the wording of questions. National research coordinators from each country review the national adaptations of individual questionnaire items and submit a report to the IEA Data Processing Center. In addition to the translation verification steps used for all TIMSS test items, a thorough item review process is used to further evaluate any items that are functioning differently in different countries according to the international item statistics. In certain cases, items have to be recoded or deleted entirely from the international database as a result of this review process.

6. CONTACT INFORMATION

For content information about the TIMSS project, contact:

Patrick Gonzales
Phone: (415) 920-9229
E-mail: patrick.gonzales@ed.gov

Mailing Address:
National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

7. METHODOLOGY AND EVALUATION REPORTS

Most of the technical documentation for TIMSS is published by the International Study Center at Boston College. The U.S. Department of Education, National Center for Education Statistics, is the source of several additional references listed below; these publications are indicated by an NCES number.

General

Gonzales, P., Calsyn, C., Jocelyn, L., Mak, K., Kastberg, D., Arafeh, S., Williams, T., and Tsen, W. (2000). *Pursuing Excellence: Comparisons of International Eighth-Grade Mathematics and Science Achievement from a U.S. Perspective, 1995 and 1999* (NCES 2001-028). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Gonzales, P., Guzmán, J.C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., and Williams, T. (2004). *Highlights From the Trends in International Mathematics and Science Study (TIMSS) 2003* (NCES 2005-005). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., and Brenwald, S. (2008). *Highlights From TIMSS 2007: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context* (NCES 2009-001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Uses of Data

Johnson, E.G. (1998). *Linking the National Assessment of Educational Progress (NAEP) and The Third International Mathematics and Science Study (TIMSS): A Technical Report (*NCES 98-499). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Johnson, E.G., and Siegendorf, A. (1998). *Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): Eighth-Grade Results (*NCES 98-500). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Kastberg, D., Roey, S., Jocelyn, L., and Williams, T. (2006). *Trends in International Mathematics and Science Study (TIMSS) 2003 U.S. Datafile and User's Guide (*NCES 2006-058). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Williams, T., Ferraro, D., Roey, S., Brenwald, S., Kastberg, D., Jocelyn, L., Smith, C., and Stearns, P. (2009). *TIMSS 2007 U.S. Technical Report and User Guide* (NCES 2009-012). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Rutkowski, L., González, E., Joncas, M., and von Davier, M. (2010). International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educational Researcher 39* (2), 142-151.

Survey Design

Martin, M.O., and Kelly, D.L. (Eds.). (1996). *TIMSS Technical Report: Volume I: Design and Development*. Boston College, International Study Center. Chestnut Hill, MA.

Martin, M.O., and Kelly, D.L. (Eds.). (1998). *TIMSS Technical Report: Volume II: Implementation and Analysis, Primary and Middle School Years*. Boston College, International Study Center. Chestnut Hill, MA.

Martin, M.O., and Kelly, D.L. (Eds.). (1999). *TIMSS Technical Report: Volume III: Implementation and Analysis, Final Year of Secondary School*. Boston College, International Study Center. Chestnut Hill, MA.

Martin, M.O., Gregory, G.D., and Stemler, S.E. (Eds.). (2000). *TIMSS 1999 Technical Report.* Boston College, International Study Center. Chestnut Hill, MA.

Martin, M.O., Mullis, I.V.S., and Chrostowski, S.J. (Eds.). (2004). *TIMSS 2003 Technical Report: Findings From IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades.* Boston College, International Study Center. Chestnut Hill, MA.

Olson, J.F., Martin, M.O., and Mullis, I.V.S. (Eds.). (2008). *TIMSS 2007 Technical Report.* Boston College International Study Center, Chestnut Hill.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Data Quality and Comparability

Martin, M.O., and Mullis, I.V.S. (Eds.). (1996). *Quality Assurance in Data Collection.* Boston College, International Study Center. Chestnut Hill, MA.

Chapter 22: Program for International Student Assessment (PISA)

1. OVERVIEW

INTERNATIONAL ASSESSMENT OF 15-YEAR-OLDS:

Assesses literacy skills in the following areas:

- Reading literacy
- Mathematics literacy
- Science literacy

The Program for International Student Assessment (PISA) is a system of international assessments that measures 15-year-old students' capabilities in reading literacy, mathematics literacy, and science literacy every 3 years. PISA 2006 was the third in this series of assessments; the fourth in the series took place in 2009. Information on PISA 2009 will not be available until December 2010, so PISA 2009 will not be included in some sections of this chapter. PISA, first implemented in 2000, was developed and is administered under the auspices of the Organization for Economic Cooperation and Development (OECD), an intergovernmental organization of industrialized countries.[1] The PISA Consortium, a group of international organizations engaged by the OECD, is responsible for coordinating the study operations across countries. The National Center for Education Statistics (NCES), in the Institute of Education Sciences at the U.S. Department of Education, is responsible for the implementation of PISA in the United States.

Purpose

PISA provides internationally comparative information on the reading, mathematics, and science literacy of students at an age that, for most jurisdictions, is near the end of compulsory schooling. The objective of PISA is to measure the "yield" of education systems, or what skills and competencies students have acquired and can apply in reading, mathematics, and science to real-world contexts by age 15. The literacy concept emphasizes the mastery of processes, the understanding of concepts, and the application of knowledge and functioning in various situations. By focusing on literacy, PISA draws not only from school curricula but also from learning that may occur outside of school.

Components

Assessment. PISA is a paper-and-pencil assessment that is designed to assess 15-year-olds' performance in reading, mathematics, and science literacy. Each student takes a 2-hour assessment. Assessment items include a combination of multiple-choice questions, closed- or short- response questions (for which answers are either correct or incorrect), and open-constructed response questions (for which answers can receive partial credit). PISA scores are reported on a scale of 0 to 1,000 with a scale mean of 500 and a scale standard deviation of 100.

Questionnaires. Students complete a 30-minute questionnaire providing information about their backgrounds, attitudes, and experiences in school. In addition, the principal of each participating school completes a 20- to 30-minute questionnaire on school characteristics and policies.

[1] Countries that participate in PISA are referred to as jurisdictions throughout this chapter.

Periodicity

PISA operates on a 3-year cycle. Each PISA assessment cycle focuses on one subject in particular, although all three subjects are assessed every year. In PISA 2000, reading literacy was the major focus. In 2003, PISA focused on mathematics literacy, and in 2006, PISA focused on science literacy. In 2009, PISA again focused on reading literacy. The remainder of this chapter focuses on the design of the 2006 administration.

2. USES OF DATA

PISA provides valuable information for comparisons of student performance across jurisdictions and over time at the national level and for some countries the subnational level. Performance in each subject area can be compared across jurisdictions in terms of:

- Jurisdictions' mean scores;
- The proportion of students in each jurisdiction reaching PISA proficiency levels;
- The scores of jurisdictions' highest performing and lowest performing students;
- The standard deviation of the distribution of scores in each jurisdiction; and
- Other measures of the distribution of performance within jurisdictions.

PISA also supports cross-jurisdictional comparisons of the performance of some subgroups of students, including students grouped by sex, immigrant status, and socioeconomic status. PISA data are not useful for comparing the performance of racial/ethnic groups across jurisdictions, because relevant racial/ethnic groups differ across jurisdictions. However, U.S. PISA datasets include information that can be used in comparing groups of students by race/ethnicity, and the poverty level of their schools within the country.

Contextual measures taken from student and principal questionnaires can be used to compare the educational contexts of 15-year-old students across jurisdictions. Caution should be taken in attempting to interpret associations between measures of educational context and student performance. The PISA assessment is intended to tap the knowledge and skills developed by students over several years as they develop factual knowledge and problem-solving skills and learn to apply them in a variety of situations. PISA contextual measures typically refer to students' current school context, which may differ from their prior school context. In the United States, data collection occurs in the fall of the school year; therefore, contextual measures may apply to only 1 or 2 months of school.

Through the collection of comparable information across jurisdictions at the student and school levels, PISA adds significantly to the knowledge base that was previously available from official national statistics.

3. KEY CONCEPTS

Literacy Types

The types of literacy measured by PISA are defined as follows (OECD 2009).

Reading literacy. An individual's capacity to understand, use, reflect on and engage with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society.

Mathematics literacy. An individual's capacity to identify and understand the role that mathematics plays in the world, make well-founded judgments, and use and engage with mathematics in ways that meet one's needs as a constructive, concerned, and reflective citizen.

Science literacy. An individual's scientific knowledge and the use of that knowledge to identify questions, acquire new knowledge, explain scientific phenomena, and draw evidence-based conclusions about science-related issues; an understanding of the characteristic features of science as a form of human knowledge and inquiry; an awareness of how science and technology shape our material, intellectual, and cultural environments; and a willingness to engage in science-related issues—and with the ideas of science—as a reflective citizen.

4. SURVEY DESIGN

The survey design for the PISA 2006 data collection is discussed in this section.

Target Population

Each jurisdiction was required to follow international standards for designing and selecting the sample, as given in the PISA sampling manual for the 2006 assessment (PISA Project Consortium 2005b). The international sampling guidelines defined the target

population and set the requirement for participation rates. The *desired* national PISA target population consisted of 15-year-old students attending educational institutions located within the jurisdiction, in 7th grade and higher. Jurisdictions were to include 15-year-old students enrolled full time in educational institutions, enrolled part time in educational institutions, enrolled in a vocational training or related type of educational program, and attending a foreign school within the jurisdiction (as well as students from other jurisdictions attending any of the programs in the first three categories). It was recognized that no testing of persons schooled in the home, workplace, or out of the jurisdiction occurred; therefore, these students were not included in the international target population.

The operational definition of an age population directly depends on the testing dates. International standards required that students in the sample be 15 years and 3 months to 16 years and 2 months at the beginning of the testing period. For PISA 2006, the testing period suggested by the OECD was between March 1, 2006, and August 31, 2006, and was required not to exceed 42 days. The United States was one of three jurisdictions to administer the assessment in fall 2006, rather than spring 2006. The United States made this choice to avoid conflicting with mandatory high-stakes testing that often occurs in the spring, based upon the PISA 2003 experience. The United States, the United Kingdom (except for Scotland), and Bulgaria moved their test date to the fall; consequently, the range of eligible birthdates in these jurisdictions was adjusted to ensure that the mean age remained consistent across all jurisdictions. In the United States, students born between July 1, 1990, and June 30, 1991, were eligible to participate in PISA 2006.

International Sample Design

In the 2006 PISA assessment, most jurisdictions used a two-stage stratified sample. The first-stage sampling units consisted of individual schools having 15-year-old students. In all but a few jurisdictions, schools were sampled systematically from a comprehensive national list of all eligible schools with probabilities that were proportional to a measure of size. This is referred to as probability proportional to size (PPS) sampling. The measure of size was a function of the estimated number of eligible 15-year-old students enrolled in the school. Prior to sampling, schools in the sampling frame were assigned to strata formed either explicitly or implicitly. The second-stage sampling units in jurisdictions using the two-stage design consisted of students within sampled schools. Once a school was selected to be in the sample, a list of the school's 15-year-old students was prepared. From each list that contained more than 35 students, 35 students were selected with equal probability, and for lists of fewer than 35 students, all students were selected. However, the minimum number of students that could be sampled within a school was 20.

Because PISA is an international survey, the types of exclusions must be defined internationally and the exclusion rates have to be limited in order to ensure that survey results are representative of the entire national school system. Thus, efforts were made to guarantee that exclusions, if they were necessary, were minimized. Exclusions could take place at the school selection stage (by excluding the whole school) or at the student selection stage.

International within-school exclusion rules for students were specified as follows:

- *Students with functional disabilities.* These were students with a moderate to severe permanent physical disability such that they could not perform in the PISA testing environment.

- *Students with intellectual disabilities.* These were students with a mental or emotional disability who had been tested as cognitively delayed or who were considered in the professional opinion of qualified staff to be cognitively delayed such that they could not perform in the PISA testing situation.

- *Students with insufficient language experience.* These were students who met the three criteria of (1) not being a native speaker in the assessment language, (2) having limited proficiency in the assessment language, and (3) having received less than a year of instruction in the assessment language. In the United States, English was the exclusive language of the assessment.

A school attended only by students who would be excluded for intellectual, functional, or linguistic reasons was considered a school-level exclusion.
School-level exclusions for inaccessibility, feasibility, or other reasons were required to cover fewer than 0.5 percent of the total number of students in the international PISA target population. International guidelines state that no more than 5 percent of a jurisdiction's desired national target population should be excluded from the sample.

A minimum of 150 schools (or all schools, if there were fewer than 150 in a participating jurisdiction) had to be selected in each jurisdiction. Within each participating school, a sample of the PISA-eligible

students was selected with equal probability. In total, a minimum sample size of 4,500 assessed students was to be achieved. If a jurisdiction had fewer than 4,500 eligible students, then the sample size was the *national defined target population*. The national defined target population included all those eligible students in the schools that were listed in the school sampling frame.

Response Rate Targets

School response rates. The PISA international guidelines for the 2006 assessment required that jurisdictions achieve an 85 percent school response rate. However, while stating that each jurisdiction must make every effort to obtain cooperation from the sampled schools, the requirements also recognized that this is not always possible. Thus, it was allowable to use substitute, or replacement, schools as a means to avoid loss of sample size associated with school nonresponse. The international guidelines stated that at least 65 percent of participating schools must be from the original sample. Jurisdictions were allowed to use replacement schools (selected during the sampling process) to increase the response rate once the 65 percent benchmark had been reached.

Each sampled school was to be assigned two replacement schools in the sampling frame. If the original sample school refused to participate, a replacement school was asked to participate. The international guidelines define the response rate as the number of participating schools (both original and replacement schools) divided by the total number of eligible original sample schools.[2]

Student response rates. A minimum response rate of 80 percent of selected students across participating schools was required. Students were deemed participants if they gave at least one response to the cognitive assessment, or if they responded to at least one student questionnaire item and either they or their parents provided the occupation of a parent or guardian.

Within each school, a student response rate of 50 percent was required for a school to be regarded as participating: the overall student response rate was computed using only students from schools with at least a 50 percent response rate. Weighted student response rates were used for assessing this standard. Each student was weighted by the reciprocal of his or her sample selection probability.

Sample Design in the United States

The design of the U.S. school sample for PISA 2006 was developed to achieve each of the international requirements set forth in the PISA sampling manual. The U.S. school sample is self-weighting, is stratified, consists of two stages (described below), and was selected using PPS sampling. The measure of size used in the first stage was the expected number of eligible 15-year-old students in the school. At the second stage, a sample of 42 students was selected from each school, regardless of size (all eligible students were selected if there were fewer than 42).

A list of schools for the U.S. sample was prepared using data from the 2003–04 Common Core of Data (CCD) and the 2003–04 Private School Universe Survey (PSS), two NCES surveys. These schools were stratified into two explicit groups: schools with large enrollments of 15-year-old students and schools with small enrollments of 15-year-old students. The frame was implicitly stratified (i.e., sorted for sampling) by five categorical stratification variables: grade span of school; control of school (public or private); region of the country; type of location relative to populous areas; and percentage of students of Black, Hispanic, and other race/ethnicities (above or below 15 percent). The last variable used for sorting within the implicit stratification was the estimated enrollment of 15-year-olds based on grade enrollments.

As in PISA 2003, schools were selected in the first stage with PPS, and students were sampled in the second stage, yielding overall equal probabilities of selection. In PISA 2000, the U.S. school sample had a three-stage design, the first of which was the selection of a sample of geographic primary sampling units (PSUs). The change to a two-stage model was made in PISA 2003 to reduce the design effects observed in the 2000 data and to minimize respondent burden on individual districts by spreading it across school districts as much as possible.

Once the school sample was drawn, it was loaded into KeyQuest, a software program written specifically for jurisdictions participating in PISA. KeyQuest was used to manage the sample, draw the student sample, track participation, and produce verification reports used to clean the data in preparation for submitting the data file.

The U.S. school sample for PISA 2006 consisted of 240 schools (from 44 states) containing at least one 7th through 12th grade class. There were 27 sampled

[2] The calculation of response rates described here is based on the formula stated in the international guidelines and is not consistent with NCES standards. A more conservative way to calculate response rates would be to include participating replacement schools in the denominator as well as in the numerator and to add replacement schools that were hard refusals to the denominator.

schools identified as ineligible or closed, reducing the sample to 209 schools.

Assessment Design

Test scope and format. In PISA 2006, the three subject domains were tested, with science as the major domain and reading and mathematics as minor domains. The development of the PISA 2006 assessment instruments was an interactive process among the PISA Project Consortium, various expert committees, and OECD members. The assessment included items submitted by participating jurisdictions and items developed by the consortium's test developers. Representatives of each jurisdiction reviewed the items for possible bias and for relevance to PISA's goals. The intention was to reflect in the assessment the national, cultural, and linguistic variety of the OECD jurisdictions. Science items were field tested in 2005 in each jurisdiction to examine their psychometric properties and identify any problematic items. Mathematics and reading items were field tested in jurisdictions that had not participated in PISA 2003. Following the field test, statistics were reviewed for each item for each jurisdiction, including percent correct, item difficulty, item discrimination, and gender differences. Items that worked differently across jurisdictions were deleted.

PISA 2006 was a paper-and-pencil assessment. Approximately one-third of the science literacy items were multiple-choice items, one-third were closed- or short-response items (for which students wrote an answer that was simply either correct or incorrect), and about one-third were open constructed-response items (for which students wrote answers that could be assigned partial credit). Items other than multiple choice were graded by trained scorers using an international scoring guide specific to each item that explicated the requirements for each score level.

Multiple-choice items were either (a) standard multiple choice, with a limited number (usually four) of responses from which students were required to select the best answer; or (b) complex multiple choice, which presented several statements, each of which required students to choose one of several possible responses (true/false, correct/incorrect, etc.). Closed- or short-response items included items which generally required students to construct a response within very limited constraints, such as mathematics items requiring a numeric answer, and items requiring a word or short phrase. Open constructed-response items required more extensive writing, or showing a calculation, and frequently included some explanation or justification. Pencils, erasers, rulers, and in some cases, calculators were provided.

Test design. The final 2006 assessment consisted of 140 science items, 48 mathematics items, and 28 reading items.

In order to cover the intended broad range of content while meeting the limit of 2 hours of individual testing time, the assessment in each domain was divided into clusters and organized into 13 booklets. Each booklet was made up of four test clusters. There were seven science clusters, four mathematics clusters, and two reading clusters. The clusters were allocated in a rotated design to the 13 booklets. The average number of items per cluster was 20 for science, 12 for mathematics, and 14 for reading. Each cluster was designed to average 30 minutes of test material.

The sampled students were randomly assigned one of the booklets, which meant each student undertook 2 hours of testing. The 2-hour test booklets were arranged in two 1-hour parts, each made up of two 30-minute time blocks. PISA's procedures provided for a short break to be taken between administrations of the two parts of the test booklet.

Every student answered science items, while mathematics and reading items were spread throughout the booklets. This assessment design was balanced so that each item cluster appeared four times, once in each of four possible locations in a booklet. Furthermore, each cluster appeared once with each other cluster. The final design, therefore, ensured that a representative sample of students responded to each cluster of items. The linked design enabled standard measurement techniques to be applied to the resulting student response data to estimate item difficulties and student abilities.

In addition to the cognitive assessment, students also received a 30-minute questionnaire designed to elicit information about their backgrounds, their attitudes, and their experiences in school. Principals in schools where PISA was administered also received a 20- to 30-minute questionnaire about their schools.

In addition to the 13 two-hour booklets, a special, optional one-hour booklet, referred to as the UH booklet (or the Une Heure booklet), was prepared for use in schools catering exclusively to students with special needs. The United States did not use the optional one-hour test booklet.

Test printing. The data collection contractor for PISA 2006, RTI International, made an error printing the test booklets in the United States, and the pagination of the booklets was consistently off by one page. The international consortium intended for the first page to

be printed on the inside of the cover; in the United States it was typically printed on the first page of plain white paper. As a result, some of the instructions in the reading section were incorrect. In some passages, students were incorrectly instructed to refer to the passage on the "opposite page" when the passage now appeared on the previous page. Because of the small number of items in the reading section, it was not possible to recalibrate the score to exclude the affected items. No incorrect page references appeared in the mathematics or science sections of the booklets.[3]

Data Collection and Processing

PISA is implemented in each jurisdiction by a National Project Manager (NPM). In the United States, the NPM works with a national data collection contractor to implement procedures prepared by the International Consortium and agreed to by the participating jurisdictions. In 2006, the U.S. national data collection contractor was RTI International.

Reference dates. The testing period suggested by the OECD was between March 1, 2006, and August 31, 2006, and was required not to exceed 42 days. However, the United States, in order to improve response rates and better accommodate school schedules, scheduled the PISA 2006 data collection from September 25 to November 22, 2006, with the agreement of the PISA Consortium. The United Kingdom (except Scotland) and Bulgaria also opted for a fall data collection period for PISA 2006.

Data collection. To implement the PISA 2006 assessment in schools, the NPMs were assisted by school coordinators and test administrators. Each NPM typically had several assistants, working from a base location (referred to as a national center). The NPM manual provided detailed information about the duties and responsibilities of the NPM. Supplementary manuals, with detailed information about particular aspects of the project, were also provided.

The test administrators were primarily responsible for administering the PISA 2006 test fairly, impartially, and uniformly, in accordance with international standards and PISA 2006 procedures. To maintain fairness, international standards stipulated that test administrators could not be the reading, mathematics, or science teacher of the students being assessed, and it was preferred that they not be a staff member at any participating school. Prior to the test date, test administrators were trained by national centers. Training included a thorough review of the test administrator manual and the script to be followed during the administration of the test and questionnaire. The PISA Project Consortium prepared a test administrator manual that described in detail the activities and responsibilities of the test administrator.

Four field supervisors and 35 test administrators were hired to work on the PISA 2006 main study in the United States. Each test administrator was assigned to one of the four field supervisors, who coordinated and monitored the test administrator's work.

The test administrator followed the instructions set forth in the international PISA test administrator manual. The students were randomly assigned one of 13 test booklets. Test administrators distributed the assessment booklets, matching the student with the preassigned booklet type according to the preprinted Student Tracking Form.

The principal at each participating school designated one person to serve as the school coordinator for PISA 2006. School coordinators were asked to work with project staff to coordinate the logistics of the test session and to ensure high student response rates. School coordinators coordinated school-related activities with the national center and the test administrators. On the test day, the school coordinator was expected to ensure that the sampled students attended the test session(s). If necessary, the school coordinator also made arrangements for a follow-up session and ensured that absent students attended the follow-up session. The PISA Project Consortium prepared a school coordinator manual that described in detail the activities and responsibilities of the school coordinator.

In the United States, schools were offered the option of conducting the assessment after school hours or on a Saturday, in addition to during school hours. This option was offered only as a refusal conversion tool and not as part of the initial recruitment materials. Of the 166 participating schools, 88 schools conducted the session during school hours, 4 conducted the session after school, and 74 participated on a Saturday. The student response rate was 91 percent during school hours and 90 percent in schools where PISA 2006 was administered after school or on a Saturday. Analyses were conducted comparing the performance of students who took the test during the regular school day with those who took the exam after school or on a Saturday. No measurable differences were found between the two groups.

Scoring. At least one-third of the PISA 2006 assessment was devoted to open constructed responses.

[3] Because of this printing error, the OECD and NCES decided not to report the PISA 2006 reading results for the United States.

The process of scoring these items was an important step in ensuring the quality and comparability of the PISA 2006 data. Detailed guidelines were developed for the scoring guides themselves, training materials to recruit scorers, and workshop materials used for the training of national scorers. Prior to the national training, the PISA Project Consortium organized training sessions to present the material and train the scoring coordinators from the participating jurisdictions, who in turn trained the national scorers.

For each test item, the scoring guide described the intent of the question and how to code the students' responses to each item. This description included the credit labels—full credit, partial credit, or no credit—attached to the possible categories of response. Also included was a system of double-digit coding for some mathematics and science items, where the first digit represented the score and the second digit represented different strategies or approaches that students used to solve the problem. The second digit generated national profiles of student strategies and misconceptions. In addition, the scoring guides included real examples of students' responses accompanied by a rationale for their classification for purposes of clarity and illustration.

To examine the consistency of this marking process in more detail within each jurisdiction (and to estimate the magnitude of the variance components associated with the use of scorers), the PISA Project Consortium conducted an interscorer reliability study on a subsample of assessment booklets. Homogeneity analysis was applied to the national sets of multiple scoring and compared with the results of the field trial. A full description of this process and the results can be found in the OECD's *PISA 2006 Technical Report* (OECD 2009).

Data Entry and Verification. The PISA Project Consortium provided participating jurisdictions with the KeyQuest data entry software. KeyQuest contained the database structures for all of the booklets, questionnaires, and tracking forms used in the main survey. Variables could be added or deleted as needed for different national options. Approved adaptations to response categories could also be accommodated. Student response data were entered directly from the test booklets and questionnaires. NPMs were responsible for ensuring that their jurisdiction's data underwent many quality checks before the data files were submitted to the PISA Project Consortium.

Once the data files were submitted to the PISA Project Consortium, they underwent two independent data cleaning procedures by data analysts. During cleaning, as many anomalies and inconsistencies as possible were identified, and through a process of extensive discussion between each national center and the PISA Project Consortium's data processing center at the Australian Council for Educational Research (ACER), an effort was made to correct and resolve all data issues.

Estimation Methods

Weighting. The use of sampling weights is necessary for the computation of statistically sound, nationally representative estimates. Survey weights adjust for the probabilities of selection for individual schools and students, for school or student nonresponse, and for errors in estimating the size of the school or the number of 15-year-olds in the school at the time of sampling.

The internationally defined weighting specifications for PISA 2006 included two base weights and five adjustments. The school base weight was defined as the reciprocal of the school's probability of selection. (For replacement schools, the school base weight was set equal to the weight of the original school it replaced.) The student base weight was given as the reciprocal of the probability of selection for each selected student from within a school.

These base weights were then adjusted for school and student nonresponse. The school nonresponse adjustment was done individually for each jurisdiction using implicit and explicit strata defined as part of the sample design. In the case of the United States, three variables were used: school control, census region, and community type. The student nonresponse adjustment was done within cells based first on students' final school nonresponse cell and their explicit stratum; within that, grade and gender were used.

All PISA 2006 analyses were conducted using these sampling weights.

Scaling. There were 13 test booklets, each containing a slightly different subset of items, in the PISA 2006 design. Each student completed one test booklet. The fact that each student completed only a subset of items means that classic test scores, such as the percent correct, are not accurate measures of student performance. Instead, scaling techniques were used to establish a common scale for all students. For PISA 2006, item response theory (IRT) was used to estimate average scores for science, mathematics, and reading literacy for each jurisdiction.

IRT identifies patterns of response and uses statistical models to predict the probability of a student

answering an item correctly as a function of his or her proficiency in answering other questions. PISA 2006 used a mixed coefficients multinomial logit IRT model. This model is similar in principle to the more familiar two-parameter logistic IRT model. With this method, the performance of a sample of students in a subject area or subarea can be summarized on a simple scale or series of scales, even when students are administered different items.

Plausible values. Scores for students are estimated as plausible values because each student completed only a subset of items. These values represent the distribution of potential scores for all students in the population with similar characteristics and identical patterns of item response. It is important to recognize that plausible values are not test scores and should not be treated as such. Plausible values are randomly drawn from the distribution of scores that could be reasonably assigned to each individual. As such, the plausible values contain random error variance components and are not optimal as scores for individuals.

The PISA 2006 student file contains many plausible values, five for each of the PISA 2006 cognitive scales (combined science literacy scale, three science literacy subscales, reading literacy scale, and mathematics literacy scale). If an analysis is to be undertaken with one of these cognitive scales, then (ideally) the analysis should be undertaken five times, once with each of the five relevant plausible value variables. The results of these five analyses are averaged; then, significance tests that adjust for variation between the five sets of results are computed.

Imputation. As with all item response scaling models, student proficiencies (or measures) are not observed; they are missing data that must be inferred from the observed item responses. There are several possible alternative approaches for making this inference. PISA uses the imputation methodology usually referred to as plausible values (described above). Plausible values are a selection of likely proficiencies for students that attained each score. Missing background data from student and principal questionnaires are not imputed for PISA reports published by NCES. In general, item response rates for variables discussed in NCES PISA reports are over the NCES standard of 85 percent.

Measuring trends. Reading literacy scales used in PISA 2000, 2003, and 2006 are directly comparable, which means that the value of 500 in PISA 2006 has the same meaning as it did in PISA 2000 and PISA 2003. However, since mathematics literacy was the major domain assessed in PISA 2003, the mathematics assessment underwent major development work and was broadened to include four domains; only two of these domains appeared in PISA 2000. As such, mathematics literacy scales are only comparable between PISA 2003 and PISA 2006. Likewise, PISA 2006 was the first major assessment of science literacy. As such, the science literacy scale in PISA 2006 is not directly comparable with those of earlier PISA assessments; however, it establishes the basis for monitoring future trends in science performance.

The PISA 2000, PISA 2003, and PISA 2006 assessments of reading, mathematics, and science are linked assessments. That is, the sets of items used to assess each domain in each year include a subset of common items. Between PISA 2000 and PISA 2003, there were 28 reading items (units and clusters), 20 mathematics items, and 25 science items that were used in both assessments. These common items are referred to as link items. The same 28 reading items were retained in 2006 to link the PISA 2006 data to PISA 2003, while 48 mathematics items from PISA 2003 were used in PISA 2006. For the science assessment, just 22 items were common to PISA 2006 and PISA 2003, and 14 were common to PISA 2006 and PISA 2000.

To establish common reporting metrics for PISA, the difficulty of the link items, measured on different occasions, is compared. Using procedures that are detailed in the *PISA 2006 Technical Report* (OECD 2009), the comparison of the item difficulties on the different occasions is used to determine a score transformation that allows the reporting of the data for a particular subject on a common scale. The change in the difficulty of each of the individual link items is used in determining the transformation; as a consequence, the sample of link items that has been chosen will influence the choice of transformation. This means that if an alternative set of link items had been chosen, the resulting transformation would be slightly different. The consequence is an uncertainty in the transformation due to the sampling of the link items, just as there is an uncertainty in values such as jurisdiction means due to the use of a sample of students.

Future Plans

After the release of PISA 2009 results in December of 2010, the next PISA assessment will be conducted in 2012. The major domain in PISA 2012 will be mathematics literacy. PISA 2012 will also include, in addition to paper-and-pencil assessments in mathematics, science, and reading literacy, computer-based assessments in mathematics and reading and a computer-based problem-solving assessment.

5. DATA QUALITY AND COMPARABILITY

A comprehensive program of continuous quality monitoring was central to ensuring full, valid implementation of the PISA 2006 procedures and the recording of deviations from these procedures. Quality monitors from the PISA Consortium visited a sample of schools in every jurisdiction to ensure that testing procedures were carried out in a consistent manner. The purpose of quality monitoring is to observe and record the implementation of the described procedures; therefore, the field operations manuals provided the foundation for all the quality monitoring procedures.

The manuals that formed the basis for the quality monitoring procedures were the NPM manual, the test administrator manual, the school coordinator manual, the school sampling preparation manual, and the PISA data management manual. In addition, the PISA data were verified at several points starting at the time of data entry.

Despite the efforts taken to minimize error, as with any study, PISA has limitations that researchers should take into consideration. Below are discussed two possible sources of error in PISA, sampling and nonsampling errors.

Sampling Error

Sampling errors occur when a discrepancy between a population characteristic and the sample estimate arises because not all members of the target population are sampled for the survey. The size of the sample relative to the population and the variability of the population characteristics both influence the magnitude of sampling error. The particular sample of 15-year-old students from fall 2006 was just one of many possible samples that could have been selected. Therefore, estimates produced from the PISA 2006 sample may differ from estimates that would have been produced had another sample of students been selected. This type of variability is called sampling error because it arises from using a sample of 15-year-old students in 2006 rather than all 15-year-old students in that year.

The standard error is a measure of the variability owing to sampling when estimating a statistic. The approach used for calculating sampling variances in PISA is Balanced Repeated Replication (BRR). This method of producing standard errors uses information about the sample design to produce more accurate standard errors than would be produced using simple random sample assumptions. Thus, the standard errors reported in PISA can be used as a measure of the precision expected from this particular sample.

Nonsampling Error

"Nonsampling error" is a term used to describe variations in the estimates that may be caused by population coverage limitations, nonresponse bias, and measurement error, as well as data collection, processing, and reporting procedures. For example, the sampling frame in the United States was limited to regular public and private schools in the 50 states and the District of Columbia and cannot be used to represent Puerto Rico or other jurisdictions. The sources of nonsampling errors are typically problems such as unit and item nonresponse, the differences in respondents' interpretations of the meaning of survey questions, response differences related to the particular time the survey was conducted, and mistakes in data preparation.

In general, it is difficult to identify and estimate either the amount of nonsampling error or the bias caused by this error. In PISA 2006, efforts were made to prevent such errors from occurring and to compensate for them when possible. For example, the design phase entailed a field test that evaluated items as well as the implementation procedures for the survey. Another potential source of nonsampling error was respondent bias, which occurs when respondents systematically misreport (intentionally or unintentionally) information in a study. One potential source of respondent bias in this survey was social desirability bias. For example, students may overstate their parents' educational attainment or occupational status. If there were no systematic differences among specific groups under study in their tendency to give socially desirable responses, then comparisons of the different groups would accurately reflect differences among groups. Readers should be aware that respondent bias may be present in this survey as in any survey. It was not possible to state precisely how such bias may affect the results.

Coverage error. Every NPM was required to define and describe their jurisdiction's national desired target population and explain how and why it might deviate from the international target population. Any hardships in accomplishing complete coverage were specified, discussed, and approved (or not) in advance. Where the national desired target population deviated from full national coverage of all eligible students, the deviations were described and enrollment data provided to measure how much that coverage was reduced. School-level and within-school exclusions from the national desired target population resulted in a national defined target population corresponding to the population of

students recorded in each jurisdiction's school sampling frame.

In PISA 2006, the United States reported 85 percent coverage of the 15-year-old population and 96 percent coverage of the national desired population. The United States reported a 4.3 percent exclusion rate, which was below the internationally acceptable exclusion rate of 5 percent.

Nonresponse error. Nonresponse error results from nonparticipation of schools and students. School nonresponse, where no replacement school participated in PISA 2006, will lead to the underrepresentation of students from the type of school that did not participate, unless weighting adjustments are made. It is also possible that only a part of the eligible population in a school (such as those 15-year-olds in a single grade) was represented by the school's student sample; this also requires weighting to compensate for the missing data from the omitted grades. Student nonresponse within participating schools occurred to varying extents. Students that could not be given achievement test scores (described in more detail below), but were not excluded for linguistic or disability reasons will be underrepresented in the data unless weighting adjustments are made.

Unit Nonresponse. In PISA 2006 in the United States, of the 240 sampled schools, 210 were eligible and 150 agreed to participate. The school response rate before replacement was 69 percent (weighted and unweighted) (Table 16). In addition to the 150 participating original schools, 20 replacement schools participated, for a total of 170 participating schools and a school response rate of 79 percent (weighted and unweighted).[4] Each of the participating schools achieved over 50 percent student participation and was included in the overall student response rate calculations.

A total of 6,800 students in the United States were sampled for the assessment. Of these students, 37 were deemed ineligible because of their enrolled grades or birthdays and 330 were deemed ineligible because they had left the school. These students were removed from the sample. Of the eligible 6,430 sampled students, an additional 250 were excluded using the criteria described earlier, for a weighted exclusion rate of 3.8 percent at the student level. Combined with the 0.5 percent of students excluded at the school level, before sampling, the overall exclusion rate for the United States was 4.3 percent. Of the 6,180 remaining sampled students, 5,620 participated. During data processing, 10 cases were deleted, leaving 5,610 cases in the final U.S. data file, for a weighted and unweighted student participation rate of 91 percent.

Table 16. U.S. weighted and unweighted school and student response rates: PISA 2006

	Response rate (percent)	
	Weighted	Unweighted
School		
Before replacement	69.0	69.4
After replacement	79.1	79.4
Student	91.0	90.8

SOURCE: Green, P., Herget, D., and Rosen, J. (2009). *User's Guide for the Program for International Student Assessment (PISA): 2006 Data Files and Database With United States Specific Variables* (NCES 2009-055). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, D.C.

Item nonresponse. For PISA 2006 in the United States, an item nonresponse bias analysis was conducted for the seven school questionnaire items with a response rate less than 85 percent and for the eight student questionnaire items with a response rate less than 85 percent. For each questionnaire item, respondents for that item were compared with nonrespondents for that item based on demographic characteristics known for everyone. These characteristics are from the CCD and PSS files, and continuous variables were made into categorical variables based on quartiles for the purpose of this analysis. For each category of each variable, bias was computed as the percentage of all item respondents who are in that category minus the percentage of all item nonrespondents who are in that category.

In PISA 2006 in the United States, five of the seven questionnaire items were significantly biased for public and private school types. There was no significant bias for any of the categories for the characteristics of total school enrollment, percent White student enrollment, and percent other student enrollment. For more details, refer to *User's Guide for the Program for International Student Assessment (PISA): 2006 Data Files and Database with United States Specific Variables* (Green, Herget, and Rosen 2009).

Measurement error. Measurement error is introduced into a survey when its test instruments do not accurately measure the knowledge or aptitude they are intended to assess.

[4] Since the U.S. school response rate was lower than the international requirement of 85 percent, the PISA Project Consortium required NCES to provide a detailed analysis of school nonresponse bias, which indicated no evidence of substantial bias resulting from school nonresponse (Green, Herget, and Rosen 2009).

Data Comparability

A number of international comparative studies already exist to measure achievement in mathematics, science, and reading, including the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). The Adult Literacy and Lifeskills Survey (ALL) was last conducted in 2003 and measured the literacy and numeracy skills of adults. A new study, the Program for the International Assessment of Adult Competencies (PIAAC), will be administered for the first time in 2011 and will assess the level and distribution of adult skills required for successful participation in the economy of participating jurisdictions. In addition, the United States has been conducting its own national surveys of student achievement for more than 35 years through the National Assessment of Educational Progress (NAEP). PISA differs from these studies in several ways.

Content. PISA is designed to measure "literacy" broadly, whereas studies such as TIMSS and NAEP have a stronger link to curriculum frameworks and seek to measure students' mastery of specific knowledge, skills, and concepts. The content of PISA is drawn from broad content areas (e.g., space and shape in mathematics) in contrast to more specific curriculum-based content, such as geometry or algebra.

Tasks. PISA also differs from other assessments in that it emphasizes the application of reading, mathematics, and science literacy to everyday situations by asking students to perform tasks that involve interpretation of real-world materials as much as possible. A study comparing the PISA, NAEP, and TIMSS mathematics assessments (Neidorf et al. 2006) found that the mathematics topics addressed by each assessment are similar, although PISA places greater emphasis on data analysis and less on algebra than does either NAEP or TIMSS. However, it is in how that content is presented that makes PISA different.

PISA uses multiple-choice items to a far lesser degree than NAEP or TIMSS, and it contains a higher proportion of items reflecting moderate to high mathematical complexity than do those two assessments. An earlier comparative analysis of the PISA, TIMSS, and NAEP mathematics and science assessments (Nohara 2001) found that compared with NAEP and TIMSS, more items in the PISA science assessment built connections to practical situations and required students to demonstrate multi-step reasoning, and fewer items used a multiple-choice format. The study also found that compared with NAEP and TIMSS, more items in the PISA mathematics assessment were set in real-life situations or scenarios, required multi-step reasoning, and required interpretation of figures and other graphical data. These tasks reflect the underlying assumption of PISA: as 15-year-olds begin to make the transition to adult life, they not only need to know how to read or use particular mathematical formulas or scientific concepts, but they also need to know how to apply this knowledge and these skills in the many different situations they will encounter in their lives.

Age-based sample. In contrast with TIMSS and PIRLS, which are grade-based assessments, PISA's sample is based on age. TIMSS assesses fourth- and eighth-graders, while PIRLS assesses fourth-graders only. The PISA sample, however, is drawn from 15-year-old students, regardless of grade level. The goal of PISA is to represent outcomes of learning rather than outcomes of schooling. By placing the emphasis on age, PISA intends to show not only what 15-year-olds have learned in school in a particular grade, but outside of school as well as over the years. PISA thus seeks to show the overall yield of an education system and the cumulative effects of all learning experience. Focusing on age 15 provides an opportunity to measure broad learning outcomes while all students are still required to be in school across the many participating jurisdictions. Finally, because years of education vary among jurisdictions, choosing an age-based sample makes comparisons across jurisdictions somewhat easier.

6. CONTACT INFORMATION

For content information about the PISA, contact:

Daniel McGrath
Phone: (202) 502-7426
E-mail: daniel.mcgrath@ed.gov

Mailing Address:
National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

7. METHODOLOGY AND EVALUATION REPORTS

Most of the technical documentation for PISA is published by the OECD. The U.S. Department of

Education, NCES, is the source of several additional references listed below.

General

Baldi, S., Jin, Y., Skemer, M., Green, P., Herget, D., and Xie, H. (2007). *Highlights From PISA 2006: Performance of U.S. 15-Year-Olds in Science and Mathematics Literacy in an International Context* (NCES 2008-016). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., Kastberg, D., and Jocelyn, L. (2004). *International Outcomes of Learning in Mathematics Literacy and Problem Solving: PISA 2003 Results From the U.S. Perspective* (NCES 2005-003). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Organization for Economic Cooperation and Development (OECD). (2005). *PISA 2003 Data Analysis Manual*. Paris: Author.

Organization for Economic Cooperation and Development (OECD). (2009). *PISA 2009 Assessment Framework - Key Competencies in Reading, Mathematics and Science*. Paris: Author.

Organization for Economic Cooperation and Development (OECD). (2007). *PISA 2006: Science Competencies for Tomorrow's World. Volume 1: Analysis.* Paris: Author.

Survey Design

Green, P., Herget, D., and Rosen, J. (2009). *User's Guide for the Program for International Student Assessment (PISA): 2006 Data Files and Database With United States Specific Variables* (NCES 2009-055). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Organization for Economic Cooperation and Development (OECD). (2009). *PISA 2006 Technical Report*. Paris: Author.

PISA Project Consortium. (2004). *Technical Standards for PISA 2006.* Retrieved November 14, 2008, from https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2006_technical_standards.pdf

PISA Project Consortium. (2005a). *Main Study National Project Manager's Manual.* Retrieved November 14, 2008, from https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2003_national_project_manager_manual.pdf

PISA Project Consortium. (2005b). *School Sampling Preparation Manual: PISA 2006 Main Study*. Retrieved November 14, 2008, from https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2003_school_sampling_manual.pdf

PISA Project Consortium. (2005c). *PISA 2006 Main Study Test Administrator's Manual.* Retrieved November 14, 2008, from https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2006_test_administrator_manual.pdf

PISA Project Consortium. (2005d). *PISA 2006 Main Study School Coordinator's Manual.* Retrieved November 14, 2008, from https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2006_school_coordinators_manual.pdf

Data Quality and Comparability

Dossey, J.A., McCrone, S.S., and O'Sullivan, C. (2006). *Problem Solving in the PISA and TIMSS 2003 Assessments* (NCES 2007-049). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Neidorf, T.S., Binkley, M., Gattis, K., and Nohara, D. (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006-029). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Nohara, D. (2001). *A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)* (NCES 2001-07). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Stephens, M., and Coleman, M.M. (2007). *Comparing PIRLS and PISA With NAEP in Reading, Mathematics, and Science*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved November 14, 2008, from

http://nces.ed.gov/Surveys/PISA/pdf/comppaper12082004.pdf

Chapter 23: International Adult Literacy Survey (IALS)

1. OVERVIEW

1994 INTERNATIONAL STUDY OF ADULT LITERACY

IALS collected:

- Background assessments
- Literacy assessments

The 1994 International Adult Literacy Survey (IALS) represented a first attempt to assess the literacy skills of entire adult populations in a framework that provided data comparable across cultures and languages. This collaborative project was designed to inform both education and labor market policy and program development activities in participating countries. The international portion of the study was carried out under the auspices of an International Steering Committee chaired by Canada, with each participating country holding a seat on the committee along with representatives from the Organization for Economic Cooperation and Development (OECD), European communities, and the United Nations Educational, Scientific and Cultural Organization.

In the United States, IALS is the fourth assessment of adult literacy funded by the federal government and conducted by the Educational Testing Service (ETS). The three previous efforts were (1) the 1992 National Adult Literacy Survey (see chapter 19); (2) the Department of Labor's (DOL) 1990 Workplace Literacy Survey; and (3) the 1985 Young Adult Literacy Assessment (funded as an adjunct to the National Assessment of Educational Progress—see chapter 18). In order to maximize the comparability of estimates across countries, IALS chose to adopt the National Adult Literacy Survey methodology and scales. Literacy was defined along three dimensions—prose, document, and quantitative. These were designed to capture an ordered set of information-processing skills and strategies that adults use to accomplish a diverse range of literacy tasks encountered in everyday life. The background data collected in IALS provide a context for understanding the ways in which various characteristics are associated with demonstrated literacy skills.

IALS was originally conducted in eight countries (Canada, Germany, Ireland, the Netherlands, Poland, Sweden, French- and German-speaking Switzerland, and the United States). A second phase was subsequently conducted in five additional countries or territories (Australia, Flemish-speaking Belgium, Great Britain, New Zealand, and Northern Ireland), and in a final phase included an additional nine countries. This chapter focuses on the first phase, in which the United States participated.

Purpose

To (1) develop scales that would permit comparisons of the literacy performance of adults (16 and older) with a wide range of abilities; (2) if such an assessment could be created, describe and compare the demonstrated literacy skills of adults in different countries.

Components

Each IALS country was given a set of model administration manuals and survey instruments as well as guidelines for adapting and translating the survey instruments. IALS instruments consisted of three parts: (1) a background questionnaire, which collected demographic information about respondents; (2) a set of core literacy tasks, which screened out respondents with very limited literacy skills; and (3) a main booklet of literacy tasks, used to calibrate literacy levels.

Background Questionnaire. The background questionnaire collected information on languages spoken or read; parents' educational attainment and employment; labor force experiences—employment status, recent labor force experiences, and occupation; reading and writing at work and looking for work; participation in adult education classes—courses taken, financial support, purpose; reading and writing in daily life (excluding work or school); family literacy—children's reading habits, the household's access to reading materials, hours spent watching television; and household information—total income and sources of income. The background questionnaire was to be administered in about 20 minutes.

Literacy Assessment—Core Literacy Tasks and Main Literacy Tasks. One hundred and fourteen tasks were grouped into three scales and divided into seven blocks (labeled A through G), which in turn were compiled into seven test booklets (numbered 1 through 7). Each booklet contained three blocks of tasks and was designed to take about 45 minutes to complete. Respondents began the cognitive part of the assessment by performing a set of six "core" tasks. Only those who were able to perform at least two of the six core tasks correctly (93 percent of respondents) were given the full assessment.

Periodicity

The first phase of data collection for IALS was conducted during the autumn of 1994 in Canada, Germany, Ireland, the Netherlands, Poland, Sweden, Switzerland (French and German-speaking cantons), and the United States. Data were collected from a second group of countries or territories—Australia, Flemish-speaking Belgium, Great Britain, New Zealand, and Northern Ireland—in 1995–96. Data were collected from a third group of countries in 1997–98. No second administration is planned.

2. USES OF DATA

IALS was designed to inform both educational and labor market policy and program development activities in participating countries. The primary objectives of the study were to

- shed light on the relationship between microeconomic variables—such as individual literacy, educational attainment, labor market participation and employment, and macroeconomic issues—such as competitiveness, growth, and restructuring;
- identify subpopulations that are economically and socially disadvantaged by their literacy skill profiles; and
- establish the comparability of assessments of adult literacy.

IALS data provide comparable information about the activities and outcomes of educational systems and institutions in participating countries. Such data can lead to improvements in accountability and policymaking. These data are relevant to policy formation due to the growing political, economic, and cultural ties between countries.

3. KEY CONCEPTS

Some of the key concepts related to the IALS literacy assessment are described below.

Literacy. The ability to use printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential.

Prose Literacy. The ability to read and use texts of varying levels of difficulty that are presented in sentence and paragraph form, including editorials, news stories, poems, and fiction.

Document Literacy. The knowledge and skills required to locate and use information contained in formats such as job applications, payroll forms, transportation schedules, maps, tables, and graphics.

Quantitative Literacy. The knowledge and skills required to apply arithmetic operations, either alone or sequentially, to numbers embedded in printed materials, such as balancing a checkbook, calculating a

tip, completing an order form, or determining the amount of interest on a loan from an advertisement.

Literacy Scales. The three scales used to report the results for prose, document, and quantitative literacy. These scales, each ranging from 0 to 500, are based on those established for the Young Adult Literacy Assessment, the DOL's Workplace Literacy Survey, and the National Adult Literacy Survey. The scores on each scale represent degrees of proficiency along that particular dimension of literacy. The scales make it possible not only to summarize the literacy proficiencies of the total population and of various subpopulations, but also to determine the relative difficulty of the literacy tasks administered in IALS.

The literacy tasks administered in IALS varied widely in terms of materials, content, and task requirements, and thus in difficulty. A careful analysis of the range of tasks along each scale provides clear evidence of an ordered set of information-processing skills and strategies along each scale. To capture this ordering, each scale was divided into five levels that reflect this progression of information-processing skills and strategies: Level 1 (0 to 225), Level 2 (226 to 275), Level 3 (276 to 325), Level 4 (326 to 375), and Level 5 (376 to 500). Level 1 comprised those adults who could consistently succeed with Level 1 literacy tasks but not with Level 2 tasks, as well as those who could not consistently succeed with Level 1 tasks and those who were not literate enough to take the test at all. Adults in Levels 2 through 4 were consistently able to succeed with tasks at their level but not with the next more difficult level of tasks. Adults in Level 5 were consistently able to succeed with Level 5 tasks. The use of three parallel literacy scales makes it possible to profile and compare the various types and levels of literacy demonstrated by adults in different countries and by subgroups within those countries.

4. SURVEY DESIGN

Statistics Canada and ETS, a private testing organization in the United States, coordinated the development and management of IALS. These organizations were assisted by national research teams from the participating countries in developing the survey design. The survey design for the 1994 IALS is described below.

Target Population

The IALS target population was the civilian, noninstitutionalized population ages 16 to 65 in each country; however, countries were also permitted to sample older adults, and several did so. All IALS samples excluded full-time members of the military and people residing in institutions such as prisons, hospitals, and psychiatric facilities.

For the United States, the target population consisted specifically of civilian noninstitutionalized residents ages 16 to 65 in the 50 states and the District of Columbia, excluding members of the armed forces on active duty, those residing outside the United States, and those with no fixed household address (i.e., the homeless or residents of institutional group quarters, such as prisons and hospitals).

Sample Design

IALS was designed to provide data representative at the national level. Each country that participated in IALS agreed to draw a probability sample that would accurately represent its civilian, noninstitutionalized population ages 16 to 65. The final IALS sample design criteria specified that each country's sample should result in at least 1,000 respondents, the minimum sample size needed to produce reliable literacy proficiency estimates. Given the different sizes of the population of persons ages 16 to 65 in the countries involved, sample sizes varied considerably from country to country (ranging from 1,500 to 8,000 per country), but sample sizes were sufficiently large in all cases to support the estimation of reliable item parameters using Item Response Theory (IRT).

IALS countries were strongly encouraged to select high-quality probability samples because the use of probability designs would make it possible to produce unbiased estimates for individual countries and to compare these estimates across the countries. Because the available data sources and resources were different in each of the participating countries, however, no single sampling methodology was imposed. Each IALS country created its own sample design. All countries used probability sampling for at least some stages of their sample designs, and some used probability sampling for all stages of sampling. Sampling designs were approved by expert review.

The sample for the United States was selected from a sample of individuals in housing units who were completing their final round of interviews for the U.S. Census Bureau's Current Population Survey (CPS) in March, April, May, and June 1994. These housing units were included in the CPS for their initial interviews in December 1992 and January, February, and March 1993. The CPS is a large-scale continuous household survey of the civilian noninstitutionalized population age 15 and over. The frame for the CPS consisted of 1990 decennial census files, which are

continually updated for new residential construction and are adjusted for undercount, births, deaths, immigration, emigration, and changes in the armed forces.

The CPS sample is selected using a stratified multistage design. Housing units that existed at the time of the 1990 population census were sampled from the census list of addresses. Housing units that did not exist at that time were sampled from lists of new construction, when available, and otherwise by area sampling methods. Occupants of housing units that came into existence between the time of the CPS sample selection and the time of the IALS fieldwork had no chance of being selected for IALS.

The IALS sample was confined to 60 of the 729 CPS primary sampling units (PSUs). Within these 60 PSUs, all persons 16 to 65 years of age in the sampled housing units were classified into 20 cells defined by race/ethnicity and education. Within each cell, persons were selected for IALS with probability proportional to their CPS weights, with the aim of producing an equal probability sample of persons within cells. A total of 4,901 persons were selected for IALS. IALS interviews were conducted in October and November 1994.

Assessment Design

The success of IALS depended on the development and standardized application of a common set of survey instruments. The test framework explicitly followed the precedent set by the National Adult Literacy Survey, basing the test on U.S. definitions of literacy along three dimensions—prose literacy, document literacy, and quantitative literacy—but extending the instruments into an international context. Study managers from each participating country were encouraged to submit materials such as news articles and documents that could be used to create tasks with the goal of building a new pool of literacy tasks that could be linked to established scales. IALS field tested 175 tasks and identified 114 that were valid across cultures. Approximately half of these tasks were based on materials from outside North America. (However, each respondent was administered only a fraction of the pool of tasks, using a variant of matrix sampling.)

Each IALS country was given a set of model administration manuals and survey instruments as well as graphic files containing the pool of IALS literacy items with instructions to modify each item by translating the English text to its own language without altering the graphic representation. Certain rules governed the item modification process. For instance, some items required respondents to perform a task that was facilitated by the use of keywords. The keyword in the question might be identical to, similar but not exactly the same as, or a synonym of the word used in the body of the item, or respondents might be asked to choose among multiple keywords in the body of the item, only one of which was correct. Countries were required to preserve these conceptual associations during the translation process. Particular conventions used in the items—for example, currency units, date formats, and decimal delimiters—were adapted as appropriate for each country.

To ensure that the adaptation process did not compromise the psychometric integrity of the items, each country's test booklets were carefully reviewed for errors of adaptation. Countries were required to correct all errors found. However, this review was imperfect in two important respects. First, it is clear that countries chose not to incorporate a number of changes that were identified during the course of the review, believing that they "knew better." Second, the availability of empirical data from the study has permitted the identification of several additional sources of task and item difficulty that were not included in the original framework, which was based on research by Irwin Kirsch of ETS and Peter Mosenthal of Syracuse University. (See 1990 publication, "Exploring Document Literacy: Variables Underlying the Performance of Young Adults," by I.S. Kirsch and P.B. Mosenthal, in *Reading Research Quarterly* 25: 5–30.) Item adaptation guidelines and item review procedures associated with subsequent rounds of IALS data collection were adapted to reflect this additional information.

The model background questionnaires contained two sets of questions: mandatory questions, which all countries were required to include; and optional questions, which were recommended but not required. Countries were not required to field literal translations of the mandatory questions, but were asked to respect the conceptual intent of each question in adapting it for use. Countries were permitted to add questions to their background questionnaires if the additional burden on respondents would not reduce response rates. Statistics Canada reviewed all background questionnaires (except Sweden's) before the pilot survey and offered comments and suggestions to each country.

Data Collection and Processing

IALS data for the first round of countries were collected through in-person household interviews in the fall of 1994. Each country mapped its national dataset into a highly structured, standardized record layout that it sent to Statistics Canada. Further description follows.

Reference dates. Respondents answered questions about jobs they may have held in the 12 months before the survey was administered.

Data collection. Statistics Canada and ETS coordinated the development and management of IALS. Participating countries were given model administration manuals and survey instruments as well as guidelines for adapting and translating the survey instruments and for handling nonresponse codings.

Countries were permitted to adapt these models to their own national data collection systems, but they were required to retain a number of key features: (1) respondents were to complete the core and main test booklets alone, in their homes, without help from another person or from a calculator; (2) respondents were not to be given monetary incentives for participating; (3) despite the prohibition on monetary incentives, interviewers were provided with procedures to maximize the number of completed background questionnaires and were to use a common set of coding specifications to deal with nonresponse. This last requirement was critical. Because noncompletion of the core and main task booklets was correlated with ability, background information about nonrespondents was needed in order to impute cognitive data for these persons.

IALS countries were instructed to obtain at least a background questionnaire from sampled individuals. All countries participating in IALS instructed interviewers to make callbacks at households that were difficult to contact.

In general, the survey was carried out in the national language. In Canada, respondents were given a choice of English or French, and in Switzerland, samples drawn from French-speaking and German-speaking cantons were required to respond in those respective languages. When respondents could not speak the designated language, attempts were made to complete the background questionnaire so that their literacy level could be estimated and the possibility of distorted results would be reduced. In the United States, the test was given in English, but a Spanish version of the background questionnaire and bilingual interviewers were available to assist individuals whose native language was not English.

Survey respondents spent approximately 20 minutes answering a common set of background questions concerning their demographic characteristics, educational experiences, labor market experiences, and literacy-related activities. Responses to these background questions made it possible to summarize the survey results using an array of descriptive variables, and also increased the accuracy of the proficiency estimates for various subpopulations. After answering the background questions, the remainder of respondents' time was spent completing a booklet of literacy tasks designed to measure their prose, document, and quantitative skills. Most of these tasks were open-ended, requiring respondents to provide a written answer.

In the United States, the IALS interview period was from October to November 1994. IALS was conducted by 149 Census Bureau interviewers. All of them had at least 5 days of interviewer training. They were given a one-day training on IALS and were provided with substantial training and reference materials based on the Canadian training package. They also performed a day of field training under the supervision of a regional office supervisor. Each interviewer had an average workload of 33 interviews, and the average number of response interviews per interviewer was 21. They were supervised by six regional supervisors who reviewed and commented on their work.

Before data collection, a letter was sent to the selected addresses describing the upcoming survey. The survey was limited to 90 minutes. If a respondent took more than 20 minutes per block, the interviewer was instructed to move the respondent on to the next block.

Data processing. As a condition of their participation in IALS, countries were required to capture and process their files using procedures that ensured logical consistency and acceptable levels of data capture error. Specifically, countries were advised to conduct complete verification of the captured scores (i.e., enter each record twice) in order to minimize error rates. One hundred percent keystroke validation was needed. Specific details about scoring are provided in a separate section below.

To create a workable comparative analysis, each IALS country was required to map its national dataset into a highly structured, standardized record layout. In addition to specifying the position, format, and length of each field, this International Record Layout included a description of each variable and indicated the categories and codes to be provided for that variable. Upon receiving a country's file, Statistics Canada performed a series of range checks to ensure compliance to the prescribed format. When anomalies were detected, countries corrected the problems and submitted new files. Statistics Canada did not, however, perform any logic or flow edits, as it was assumed that participating countries performed this step themselves.

Editing. Most countries followed IALS guidelines, verifying 100 percent of their data capture operation. The two countries that did not comply with this recommendation conducted sample verifications, one country at 20 percent and the other at 10 percent. Each country coded and edited its own data, mapping its national dataset into the detailed International Record Layout, which included a description of each variable and indicated the categories and codes to be provided for that variable. Industry, occupation, and education were coded using the standard international coding schemes: the International Standard Industrial Classification (ISIC), the International Standard Classification of Occupations (ISCO), and the International Standard Classification of Education (ISCED). Coding schemes were provided for open-ended items; the coding schemes came with specific instructions so that coding error could be contained to acceptable levels.

Scoring. Respondents' literacy proficiencies were estimated based on their performance on the cognitive tasks administered in the assessment. Because the open-ended items used in IALS elicited a large variety of responses, responses had to be grouped in order to summarize the performance results. As they were scored, responses to IALS open-ended items were classified as correct, incorrect, or omitted. The models employed to estimate ability and difficulty were predicated on the assumption that the scoring rubrics developed for the assessment were applied in a consistent fashion within and between countries. To reinforce the importance of consistent scoring, a meeting of national study managers and chief scorers was held prior to the commencement of scoring for the main study. The group spent 2 days reviewing the scoring rubrics for all the survey items. Where this review uncovered ambiguities and situations not covered by the guides, clarifications were agreed to collectively, and these clarifications were then incorporated into the final rubrics. To provide ongoing support during the scoring process, Statistics Canada and ETS maintained a joint scoring hotline. Any scoring problems encountered by chief scorers were resolved by this group, and decisions were forwarded to all national study managers. Study managers conducted intensive scoring training using the scoring manual and discussed unusual responses with scorers. They also offered additional training to some scorers, as needed, to raise their accuracy to the level achieved by other scorers.

To maintain coding quality within acceptable levels of error, each country undertook to rescore a minimum of 10 percent of all assessments. Where significant problems were encountered, larger samples of a particular scorer's work were to be reviewed and, where necessary, their entire assignments rescored. Countries were not required to resolve contradictory scores in the main survey (as they had been in the pilot), since outgoing agreement rates were far above minimum acceptable tolerances.

Since there could still be significant differences in the consistency of scoring between countries, countries agreed to exchange at least 300 randomly selected booklets with another country sharing the same test language. In all cases where serious discrepancies were identified, countries were required to rescore entire items or discrepant code pairs.

Intra-country rescoring. A variable sampling ratio procedure was set up to monitor scoring accuracy. At the beginning of scoring, almost all responses were rescored to identify inaccurate scorers and to detect unique or difficult responses that were not covered in the scoring manual. After a satisfactory level of accuracy was achieved, the rescoring ratio was dropped to a maintenance level to monitor the accuracy of all scorers. Average agreements were calculated across all items. Precautions were taken to ensure that the first and second scores were truly independent.

Intercountry rescoring. To determine intercountry scoring reliabilities for each item, the responses of a subset of examinees were scored by two separate groups. Usually, these scoring groups were from different countries. Intercountry score reliabilities were calculated by Statistics Canada, and then evaluated by ETS. Based on the evaluation, every country was required to introduce a few minor changes in scoring procedures. In some cases, ambiguous instructions in the scoring manual were found to be causing erroneous interpretations and therefore lower reliabilities.

Using the intercountry score reliabilities, researchers could identify poorly constructed items, ambiguous scoring criteria, erroneous translations of items or scoring criteria, erroneous printing of items or scoring criteria, scorer inaccuracies, and, most important, situations in which one country consistently scored differently from another. In the latter circumstance, scorers in one country may consistently rate a certain response as being correct while those in another country score the same response as incorrect. ETS and Statistics Canada examined scoring carefully to identify situations in which scorers in one country were consistently rating a certain response as being correct while those in another country were scoring the same response as incorrect. Where a systematic error was identified in a particular country, the original scores for that item were corrected for the entire sample.

Estimation Methods

Weighting was used in the 1994 IALS to adjust for sampling and nonresponse. Responses to the literacy tasks were scored using IRT scaling. A multiple imputation procedure based on plausible values methodology was used to estimate the literacy proficiencies of individuals who completed literacy tasks.

Weighting. IALS countries used different methods for weighting their samples. Countries with known probabilities of selection could calculate a base weight using the probability of selection. To adjust for unit nonresponse, all countries poststratified their data to known population counts, and a comparison of the distribution of the age and sex characteristics of the actual and weighted samples indicates that the samples were comparable to the overall populations of IALS countries. Another commonly used approach was to weight survey data to adjust the rough estimates produced by the sample to match known population counts from sources external to IALS. This "benchmarking" procedure assumes that the characteristics of nonrespondents are similar to those of respondents. It is most effective when the variables used for benchmarking are strongly correlated with the characteristic of interest—in this case, literacy levels. For IALS, the key benchmarking variables were age, employment status, and education. All of the IALS countries benchmarked to at least one of these variables. The United States used education.

Weights for the U.S. IALS sample included two components. The first assigned weights to CPS respondents, and the second assigned weights to IALS respondents.

The CPS weighting scheme was a complex one involving three components: basic weighting, noninterview adjustment, and ratio adjustment. The basic weighting compensated for unequal selection probabilities. The noninterview adjustment compensated for nonresponse within weighting cells created by clusters of PSUs of similar size; Metropolitan Statistical Area (MSA) clusters were subdivided into central city areas, and the balance of the MSA and non-MSA clusters were divided into urban and rural areas. The ratio adjustment made the weighted sample distributions conform to known distributions on such characteristics as age, race, Hispanic origin, sex, and residence.

The weights of persons sampled for IALS were adjusted to compensate for the use of the four rotation groups, the sampling of the 60 PSUs, and the sampling of persons within the 60 PSUs. The IALS noninterview adjustment compensated for sampled persons for whom no information was obtained because they were absent, refused to participate, had a short-term illness, had moved, or had experienced an unusual circumstance that prevented them from being interviewed. Finally, the IALS ratio adjustment ensured that the weighted sample distributions across a number of education groups conformed to March 1994 CPS estimates of these numbers.

Scaling. The scaling model used in IALS was the two-parameter logistic model based on IRT.

Items developed for IALS were based on the framework used in three previous large-scale assessments: the Young Adult Literacy Assessment, the DOL survey, and the National Adult Literacy Survey. As a result, IALS items shared the same characteristics as the items in these earlier surveys. The English versions of IALS items were reviewed and tested to determine whether they fit into the literacy scales in accordance with the theory and whether they were consistent with the National Adult Literacy Survey data. Quality control procedures for item translation, scoring, and scaling followed the same procedures used in the National Adult Literacy Survey and extended the methods used in other international studies.

Identical item calibration procedures were carried out separately for each of the three literacy scales: prose, document, and quantitative literacy. Using a modified version of Mislevy and Bock's 1982 BILOG computer program—see *BILOG: Item analysis and test scoring with binary logistic models*, Scientific Software—the two-parameter logistic IRT model was fit to each item using sample weights. BILOG procedures are based on an extension of the marginal-maximum-likelihood approach described by Bock and Aitkin in their 1981 *Psychometrika* article, "Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm."

Most of the items administered in IALS were successful from a psychometric standpoint. However, despite stringent efforts at quality control, some of the assessment items did not meet the criteria for inclusion in the final tabulation of results. Specifically, in carrying out the IRT modeling used to create the three literacy scales, researchers found that a number of assessment items had significantly different item parameters across IALS countries.

Imputation. A respondent had to complete the background questionnaire, pass the core block of literacy tasks, and attempt at least five tasks per literacy scale

in order for researchers to be able to estimate his or her literacy skills directly. Literacy proficiency data were imputed for individuals who failed or refused to perform the core literacy tasks and for those who passed the core block but did not attempt at least five tasks per literacy scale. Because the model used to impute literacy estimates for nonrespondents relied on a full set of responses to the background questions, IALS countries were instructed to obtain at least a background questionnaire from sampled individuals. IALS countries were also given a detailed nonresponse classification to use in the survey.

Literacy proficiencies of respondents were estimated using a multiple imputation procedure based on plausible values methodology. Special procedures were used to impute missing cognitive data.

Literary proficiency estimation (plausible values). A multiple imputation procedure based on plausible values methodology was used to estimate respondents' literacy proficiency in the 1994 IALS. When a sampled individual decided to stop the assessment, the interviewer used a standardized nonresponse coding procedure to record the reason why the person was stopping. This information was used to classify nonrespondents into two groups: (1) those who stopped the assessment for literacy-related reasons (e.g., language difficulty, mental disability, or reading difficulty not related to a physical disability); and (2) those who stopped for reasons unrelated to literacy (e.g., physical disability or refusal). About 45 percent of the individuals did not complete the assessment for reasons related to their literacy skills; the other respondents gave no reason for stopping or gave reasons unrelated to their literacy.

When individuals cited a literacy-related reason for not completing the cognitive items, it implies that they were unable to respond to the items. On the other hand, citing reasons unrelated to literacy implies nothing about a person's literacy proficiency. Based on these interpretations, IALS adapted a procedure originally developed for the National Adult Literacy Survey to treat cases in which an individual responded to fewer than five items per literacy scale, as follows: (1) if the individual cited a literacy-related reason for not completing the assessment, then all consecutively missing responses at the end of the block of items were treated as wrong; and (2) if the individual cited reasons unrelated to literacy for not completing the assessment, then all consecutively missing responses at the end of a block were treated as "not reached."

Proficiency values were estimated based on respondents' answers to the background questions and the cognitive items. As an intermediate step, the functional relationship between these two sets of information was calculated, and this function was used to obtain unbiased proficiency estimates with reduced error variance. A respondent's proficiency was calculated from a posterior distribution that was the multiple of two functions: a conditional distribution of proficiency, given responses to the background questions; and a likelihood function of proficiency, given responses to the cognitive items.

Recent Changes

Since IALS was a one-time assessment, there are no changes to report.

Future Plans

There are no plans to conduct IALS again. However, a new survey, the Adult Literacy and Lifeskills Survey (ALL), was administered in 2003 (see chapter 24). The aspects of this survey that address literacy were built on methodologies used in IALS.

5. DATA QUALITY AND COMPARABILITY

The literacy tasks contained in IALS and the adults asked to participate in the survey were samples drawn from their respective universes. As such, they were subject to some measurable degree of uncertainty. IALS implemented procedures to minimize both sampling and nonsampling errors. The IALS sampling design and weighting procedures assured that participants' responses could be generalized to the population of interest. Scientific procedures employed in the study design and the scaling of literacy tasks permitted a high degree of confidence in the resulting estimates of task difficulty. Quality control activities continued during interviewer training, data collection, and processing of the survey data.

In addition, special evaluation studies were conducted to examine issues related to the quality of IALS. These studies included (1) an external evaluation of IALS methodology; (2) an examination of how similar or different the sampled persons were from the overall population; (3) an evaluation of the extent to which the literacy levels of the population in the database for each nation were predictable based on demographic characteristics; (4) an examination of the assumption of unidimensionality; and (5) an evaluation of the construct validity of the adult literacy scales.

Sampling Error

Because IALS employed probability sampling, the results were subject to sampling error. Although small, this error was higher in IALS than in most studies because the cost of surveying adults in their homes is so high. Most countries simply could not afford large sample sizes.

Each country provided a set of replicate weights for use in a jackknife variance estimation procedure.

There were three situations in which nonprobability-based sampling methods were used: France and Germany used "random route" procedures for selecting households into their samples, and Switzerland used an alphabetic sort to select one member of each household. However, based on the available evidence, it is not believed that these practices introduced significant bias into the survey estimates.

In 1998, the U.K. Office of National Statistics coordinated the European Adult Literacy Review, a split-sample survey intended, in part, to measure the effects of sampling methods on the IALS results. This follow-up survey compared an IALS sample design with an alternative, standardized "best practice" design. Although certain differences were noted between the two samples, the IALS sample design was not confirmed to be inferior to the "best practice" design.

Nonsampling Error

The key sources of nonsampling error in the 1994 IALS were differential coverage across countries and nonresponse bias, which occurred when different groups of sampled individuals failed to participate in the survey. Other potential sources of nonsampling error included deviations from prescribed data collection procedures and errors of logic that resulted from mapping idiosyncratic national data into a rigid international format. Scoring error, associated with scoring open-ended tasks reliably within and between countries, also occurred. Finally, because IALS data were collected and processed independently by the various countries, the study was subject to uneven levels of commonplace data capture, data processing, and coding errors.

Three studies were conducted to examine the possibility of nonresponse bias. Because the sampling frames for Canada and the United States contained information about the characteristics of sampled individuals, it was possible to compare the characteristics of respondents and nonrespondents, particularly with respect to literacy skill profiles. The Swedish National Study Team also commissioned a nonresponse follow-up study.

Coverage error. The design specifications for IALS stated that in each country the study should cover the civilian, noninstitutionalized population ages 16 to 65. It is the usual practice to exclude the institutional population from national surveys because of the difficulties in conducting interviews in institutional settings. Similarly, it is not uncommon to exclude certain other parts of a country's population that pose difficult survey problems (e.g., persons living in sparsely populated areas). The intended coverage of the surveys generally conformed well to the design specifications: each of the IALS countries attained a high level of population coverage, ranging from a low of 89 percent in Switzerland to a high of 99 percent in the Netherlands and Poland. However, it should be noted that actual coverage is generally lower than the intended coverage because of deficiencies in sampling frames and sampling frame construction (e.g., failures to list some households and some adults within listed households). In the United States, for example, comparing population sizes estimated from the survey with external benchmark figures suggests that the overall coverage rate for the CPS (the survey from which the IALS sample was selected) is about 93 percent, but that it is much lower for certain population subgroups (particularly young Black male adults).

Nonresponse error. For IALS, several procedures were developed to reduce biases due to nonresponse, based on how much of the survey the respondent completed.

Unit nonresponse. The definition of a respondent for IALS was a person who partially or fully completed the background questionnaire. Unweighted response rates varied considerably from country to country, ranging from a high of 69 percent (Canada, Germany) to a low of 45 percent (the Netherlands), with four countries in the 55–60 percent range.

In the United States, which had a response rate of 60 percent, nonresponse to IALS occurred for two reasons: (1) some individuals did not respond to the CPS; and (2) some of the CPS respondents selected for IALS did not respond to the IALS instruments. In any given month, nonresponse to the CPS is typically quite low, around 4 to 5 percent. Its magnitude in the expiring rotation groups employed for IALS selection is not known. About half of the CPS nonresponse is caused by refusals to participate, while the remainder is caused by temporary absences, other failures to contact individuals, the inability of individuals contacted to respond, and unavailability for other reasons.

A sizable proportion of the nonresponse to the IALS background questionnaire was attributable to persons who had moved. For budgetary reasons, it was decided

that persons who were not living at the CPS addresses at the time of the IALS interviews would not be contacted. This decision had a notable effect on the sample of students, who are sampled in dormitories and other housing units in the CPS only if they do not officially reside at their parents' homes. Those who reside at their parents' homes are included in the CPS at that address, but because most of these students were away at college during the IALS interview period (October to November 1994), they could not respond to IALS.

The high level of nonresponse for college students could cause a downward bias in the literacy skill-level estimates. This group represents only a small proportion of the U.S. population, however, so the potential bias is likely to be quite small. Furthermore, a comparison of IALS results to the U.S. National Adult Literacy Survey data discounts this as a major source of bias.

Item nonresponse. The weighted percentage of omitted responses for the U.S. IALS sample ranged from 0 to 18 percent.

Not-reached responses were classified into two groups: nonparticipation immediately or shortly after the background information was collected; and premature withdrawal from the assessment after a few cognitive items were attempted. The first type of not-reached response varied a great deal across countries according to the frames from which the samples were selected. The second type of not-reached response was due to quitting the assessment early, resulting in incomplete cognitive data. Not-reached items were treated as if they provided no information about the respondent's proficiency, so they were not included in the calculation of likelihood functions for individual respondents. Therefore, not-reached responses had no direct impact on the proficiency estimation for subpopulations. The impact of not-reached responses on the proficiency distributions was mediated through the subpopulation weights.

Measurement error. Assessment tasks were selected to ensure that, among population subgroups, each literacy domain (prose, document, and quantitative) was well covered in terms of difficulty, stimuli type, and content domain. The IALS item pool was developed collectively by participating countries. Items were subjected to a detailed expert analysis at ETS and vetted by participating countries to ensure that the items were culturally appropriate and broadly representative of the population being tested. For each country, experts who were fluent in both English and the language of the test reviewed the items and identified ones that had been improperly adapted. Countries were asked to correct problems detected during this review process. To ensure that all of the final survey items had a high probability of functioning well, and to familiarize participants with the unusual operational requirements involved in data collection, each country was required to conduct a pilot survey. Although the pilot surveys were small and typically were not based strictly on probability samples, the information they generated enabled ETS to reject items, to suggest modifications to a few items, and to choose good items for the final assessment. ETS's analysis of the pilot survey data and recommendations for the final test design were presented to and approved by participating countries.

Data Comparability

While most countries closely followed the data collection guidelines provided, some did deviate from the instructions. First, two countries (Sweden and Germany) offered participation incentives to individuals sampled for their survey. The incentive paid was trivial, however, and it is unlikely that this practice distorted the data. Second, the doorstep introduction provided to respondents differed somewhat from country to country. Three countries (Germany, Switzerland, and Poland) presented the literacy test booklets as a review of the quality of published documents rather than as an assessment of the respondent's literacy skills. A review of these practices suggested that they were intended to reduce response bias and were warranted by cultural differences in respondents' attitudes toward being tested. Third, there were differences across the countries in the way in which interviewers were paid. No guidelines were provided on this subject, and the study teams therefore decided what would work best in their respective countries. Fourth, several countries adopted field procedures that undermined the objective of obtaining completed background questionnaires for an overwhelming majority of selected respondents.

This project was designed to produce data comparable across cultures and languages. After one of the countries in the first round raised concerns about the international comparability of the survey data, Statistics Canada decided that the IALS methodology should be subjected to an external evaluation. In the judgment of the expert reviewers, the considerable efforts that were made to develop standardized survey instruments for the different nations and languages were successful, and the data obtained from them should be broadly comparable.

However, the standardization of procedures with regard to other aspects of survey methodology was not

achieved to the extent desired, resulting in several weaknesses. Nonresponse proved to be a particular weakness, with generally very high nonresponse rates and variation in nonresponse adjustment procedures across countries. For some countries the sample design was problematic, resulting in some unknown biases. The data collection and its supervision differed between participating countries, and some clear weaknesses were evident for some countries. The reviewers felt that the variation in survey execution across countries was so large that they recommended against publication of comparisons of overall national literacy levels. They did, however, despite the methodological weaknesses, recommend that the survey results be published. They felt that the instruments developed for measuring adult literacy constituted an important advance, and the results obtained for the instruments in the first round of IALS were a valuable contribution to the field. They recommended that the survey report focus on analyses of the correlates of literacy (e.g., education, occupation, and age) and the comparison of these correlates across countries. Although these analyses might also be distorted by methodological problems, they believed that the analyses were likely to be less affected by these problems than were the overall literacy levels.

6. CONTACT INFORMATION

For content information on IALS, contact:

Eugene Owen
Phone: (202) 502-7422
E-mail: eugene.owen@ed.gov

Mailing Address:
National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

7. METHODOLOGY AND EVALUATION REPORTS

Murray, T.S., Kirsch, I.S. and Jenkins, L.B. (eds.). (1997). *Adult Literacy in OECD Countries: Technical Report on the First International Adult Literacy Survey* (NCES 98-053). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Chapter 24: Adult Literacy and Lifeskills Survey (ALL)

1. OVERVIEW

ADULT LITERACY AND LIFESKILLS SURVEY

ALL collected:

- Background assessments
- Literacy assessments in prose literacy, document literacy, numeracy, and problem-solving domains

The Adult Literacy and Lifeskills Survey (ALL) is an international comparative study designed to provide participating countries, including the United States, with information about the skills of their adult populations ages 16 to 65. The development and management of the study were coordinated by Statistics Canada and the Educational Testing Service (ETS) in collaboration with the National Center for Education Statistics (NCES) of the U.S. Department of Education; the Organization for Economic Cooperation and Development (OECD); the Regional Office of Education for Latin America and the Caribbean (OREALC); and the Institute for Statistics (UIS) of the United Nations Educational, Scientific, and Cultural Organization (UNESCO).

ALL measured the literacy and numeracy skills of a nationally representative sample from each participating country. On a pilot basis, ALL also measured adults' problem-solving skills and gathered information on their familiarity with information and communication technology (ICT). ALL builds on the foundation of earlier studies of adult literacy. Chief among these earlier studies is the International Adult Literacy Survey (IALS), which was conducted in three phases (1994, 1996, and 1998) in 20 nations, including the United States. The following six countries participated in ALL: Italy, Norway, Switzerland, Bermuda, Canada, and the United States.

Purpose

To (1) profile and compare the literacy skills in adult populations; (2) profile and compare the level and distribution of directly assessed numeracy skills among adult populations in participating countries; (3) profile and compare the level and distribution of problem-solving skills among the adult populations of the countries surveyed; and (4) collect comparable data on participation in formal adult education.

Components

Each ALL country was given a set of model administration manuals and survey instruments as well as guidelines for adapting and translating the survey instruments. ALL instruments consisted of three parts: (1) a background questionnaire, which collected demographic information about respondents; (2) a set of core literacy tasks, which screened out respondents with very limited literacy skills; and (3) a main booklet of literacy tasks, used to calibrate literacy levels.

Background Questionnaire. The background questionnaire collected general participant information (such as sex, age, race/ethnicity, education level, and labor force status) and posed more targeted questions related to literacy practices, familiarity with ICT, education coursetaking, and health.

Literacy Assessment.
Core literacy tasks. The core literacy tasks were presented to respondents once they had completed the background questionnaire. The booklet for the core literacy tasks contained six simple tasks. Only those who answered at least two of the core tasks correctly were given the full assessment.

Main literacy tasks. The main literacy tasks for the ALL psychometric assessment consisted of tasks in prose literacy, document literacy, numeracy, and problem-solving domains. The assessment included four 30-minute blocks of literacy items (i.e., prose and document literacy), two 30-minutes blocks of numeracy items, and two 30-minute blocks of problem-solving items. A four-domain ALL assessment was implemented in Bermuda, Canada, Italy, Norway, and the French- and German-language regions of Switzerland. The United States and the Italian-language region of Switzerland carried out a three-domain ALL assessment that excluded the problem-solving domain. The blocks of assessment items were organized into 28 task booklets in the case of the four-domain assessment and into 18 task booklets for the three-domain assessment. The assessment blocks were distributed to the task booklets according to a balanced incomplete block (BIB) design whereby each task booklet contained two blocks of items.

Periodicity

ALL was conducted between the fall of 2003 and early spring 2004. In the United States, data collection for the main study took place between January and June 2003.

2. USES OF DATA

ALL sought to provide researchers with information on skill gain and loss in the adult population. This was achieved through the measurement of prose and document literacy. Furthermore, the study extended the range of skills measured by adding tasks for problem-solving, numeracy, and ICT skills. This allows researchers to examine the profiles of important foundation skills. The study makes it possible to explore the interrelationships among skill domains as well as their links to major antecedents and outcomes, such as the quantity and quality of initial education and the impact of skills on employability, wages, and health.

In addition, information from ALL addresses questions such as the following:

- What is the distribution of literacy and numeracy skills among American adults? How do these skill distributions compare to those of other countries?
- What is the relationship between these literacy skills and the economic, social, and personal characteristics of individuals? For example: Do different age or linguistic groups manifest different skill levels? Do males and females perform differently? At what kinds of jobs do people at various literacy levels work? What wages do they earn? How do adults who have completed different levels of education perform?
- What is the relationship between these skills and the economic and social characteristics of nations? For example, how do the skills of the adult labor force of a country match up with areas of the economy that are growing?

3. KEY CONCEPTS

Four skill domains are conceptualized in ALL: prose literacy, document literacy, numeracy, and problem solving. Two of them, namely, prose and document literacy, are defined and measured in the same manner as in IALS (see chapter 23). Numeracy and problem solving are new domains.

Prose literacy. The knowledge and skills needed to understand and use information from texts, including editorials, news stories, brochures, and instruction manuals.

Document literacy. The knowledge and skills required to locate and use information contained in various formats, including job applications, payroll forms, transportation schedules, maps, tables, and charts.

Numeracy. The knowledge and skills required to effectively manage the mathematical demands of diverse situations.

Problem solving. Problem solving involves goal-directed thinking and action in situations for which no routine solution procedure is available. The problem solver has a more or less well-defined goal, but does not immediately know how to reach it. The incongruence of goals and admissible operators constitutes a problem. The understanding of the problem situation and its step-by-step transformation

based on planning and reasoning constitute the process of problem solving.

Literacy scale. For each skill assessment domain, proficiency is denoted on a scale ranging from 0 to 500 points. Each score denotes a point at which a person has an 80 percent chance of successfully completing tasks that are associated with a similar level of difficulty. For the prose and document literacy domains as well as the numeracy domain, experts defined five broad levels of difficulty, each corresponding to a range of scores. For the problem-solving domain, experts defined four broad levels of difficulty.

4. SURVEY DESIGN

Each participating country was required to design and implement the Adult Literacy and Lifeskills Survey according to specified guidelines and standards. These ALL standards established the minimum survey design and implementation requirements for the following project areas: survey planning; target population; method of data collection; sample frame; sample design; sample selection; literacy assessment design; background questionnaire; task booklets; instrument requirements to facilitate data processing; data collection; respondent contact strategy; response rate strategy; interviewer hiring, training, and supervision; data capture; coding and scoring; data file format and editing; weighting; estimation; confidentiality; survey documentation; and pilot survey.

Target Population

Each participating country designed a sample to be representative of its civilian noninstitutionalized population ages 16 to 65 (inclusive). Countries were also at liberty to include adults over the age of 65 in the sample provided that a minimum suggested sample size requirement was satisfied for the 16 to 65 age group. Canada opted to include in its target population adults over the age of 65. All of the remaining countries restricted the target population to the 16 to 65 age group. Exclusions from the target population for practical operational reasons were acceptable provided a country's survey population did not differ from the target population by more than 5 percent (i.e., provided that the total number of exclusions from the target population due to undercoverage was not more than 5 percent of the target population). All countries indicate that this 5 percent requirement was satisfied. Each country chose or developed a sample frame to cover the target population.

Sample Design

Each participating country was required to use a probability sample representative of the national population ages 16 to 65. A sample size of 5,400 completed cases in each official language was recommended for each country that was implementing the full ALL psychometric assessment (i.e., comprising the prose literacy, document literacy, numeracy, and problem-solving domains). A sample size of 3,420 complete cases in each official language was recommended if the problem-solving domain was excluded from the ALL assessment.

The available sampling frames and resources varied from one country to another. Therefore, the particular probability sample design to be used was left to the discretion of each country. Each country's proposed sample design was reviewed by Statistics Canada to ensure that the sample design standards and guidelines were satisfied.

A stratified multistage probability sample design was employed in the United States. The first stage of sampling consisted of selecting a sample of 60 primary sampling units (PSUs) from a total of 1,880 PSUs that were formed using a single county or a group of contiguous counties, depending on the population size and the area covered by a county or counties. The PSUs were stratified on the basis of the social and economic characteristics of the population, as reported in the 2000 census. The following characteristics were used to stratify the PSUs: region of the country, Metropolitan Statistical Area (MSA), population size, percentage of African-American residents, percentage of Hispanic residents, and per capita income. The largest PSUs in terms of population size were included in the sample with certainty. For the remaining PSUs, one PSU per stratum was selected with probability proportional to the population size.

At the second sampling stage, a total of 505 geographic segments were systematically selected with probability proportional to population size from the sampled PSUs. Segments consist of area blocks (as defined by the 2000 census) or combinations of two or more nearby blocks. They were formed to satisfy criteria based on population size and geographic proximity. The third stage of sampling involved the listing of the dwellings in the selected segments and the subsequent selection of a random sample of dwellings. An equal number of dwellings was selected from each sampled segment. At the fourth and final stage of sampling, one eligible person was randomly selected within households with fewer than four eligible adults. In households with four or more eligible persons, two adults were randomly selected.

Assessment Design

A BIB assessment design was used to measure the skill domains. The BIB design comprised a set of assessment tasks organized into smaller sets of tasks, or blocks. Each block contained assessment items from one of the skill domains and covered a wide range of difficulty (i.e., from easy to difficult). The blocks of items were organized into task booklets according to a BIB design. Individual respondents were not required to take the entire set of tasks. Instead, each respondent was randomly administered one of the task booklets.

ALL assessment. The ALL psychometric assessment consisted of the prose literacy, document literacy, numeracy, and problem-solving domains. The assessment included four 30-minute blocks of literacy items (i.e., prose and document literacy), two 30-minute blocks of numeracy items, and two 30-minute blocks of problem-solving items. A four-domain ALL assessment was implemented in Bermuda, Canada, Italy, Norway, and the French- and German-language regions of Switzerland. The United States and the Italian-language region of Switzerland carried out a three-domain ALL assessment that excluded the problem-solving domain.

The blocks of assessment items were organized into 28 task booklets in the four-domain assessment and into 18 task booklets in the three-domain assessment. The assessment blocks were distributed to the task booklets according to a BIB design whereby each task booklet contained two blocks of items. The task booklets were randomly distributed among the selected sample. In addition, the data collection activity was closely monitored in order to obtain approximately the same number of complete cases for each task booklet, except for two-task booklets in the three-domain assessment containing only numeracy items, which required a larger number of complete cases.

Data Collection and Processing

The data collection for the ALL project took place between the fall of 2003 and early spring 2004, depending on the country. However, in the United States, data collection for the main study took place between January and June 2003. In the United States, a nationally representative sample of 3,420 adults ages 16 to 65 participated in ALL. Trained interviewers administered approximately 45 minutes of background questions and 60 minutes of assessment items to participants in their homes.

Reference dates. Respondents answered questions about jobs they may have held in the 12 months before the survey was administered.

Data collection. The ALL survey design combined educational testing techniques with those of household survey research to measure literacy and provide the information necessary to make these measures meaningful. The respondents were first asked a series of questions to obtain background and demographic information on educational attainment, literacy practices at home and at work, labor force information, ICT use, adult education participation, and literacy self-assessment. Once the background questionnaire had been completed, the interviewer presented a booklet containing six simple tasks (the core tasks). Respondents who passed the core tasks were given a much larger variety of tasks, drawn from a pool of items grouped into blocks; each booklet contained two blocks that represented about 45 items. No time limit was imposed on respondents, and they were urged to try each item in their booklet. Respondents were given the maximum leeway to demonstrate their skill levels, even if their measured skills were minimal.

To ensure high-quality data, ALL guidelines specified that each country should work with a reputable data collection agency or firm, preferably one with its own professional, experienced interviewers. The interviews were to be conducted in the home in a neutral, nonpressured manner. Interviewer training and supervision was to be provided that emphasized the selection of one person per household (if applicable), the selection of one of the 28 main task booklets (if applicable), the scoring of the core task booklet, and the assignment of status codes. Finally, the interviewers' work was to be supervised by the use of quality checks—frequent quality checks at the beginning of the data collection and fewer quality checks throughout the remainder of the data collection—and by having help available to interviewers during entire the data collection period.

Several precautions were taken against nonresponse bias. Interviewers were specifically instructed to return several times to nonrespondent households in order to obtain as many responses as possible. In addition, all countries were asked to ensure that the address information provided to interviewers was as complete as possible in order to reduce potential household identification problems. Countries were asked to complete a debriefing questionnaire after the study in order to demonstrate that the guidelines had been followed, as well as to identify any collection problems they had encountered.

The United States administered the survey only in English. It used 106 interviewers during the data collection process, assigning approximately 64 cases to each interviewer. Professional interviewers were used

to conduct the survey, although approximately one-quarter of the interviewers had no previous survey experience.

Data processing. As a condition of their participation in ALL, countries were required to capture and process their files using procedures that ensured logical consistency and acceptable levels of data capture error. Specifically, countries were advised to conduct complete verification of the captured scores (i.e., enter each record twice) in order to minimize error rates. Because the process of accurately capturing the task scores is essential to high data quality, 100 percent keystroke verification was required.

Each country was also responsible for coding industry, occupation, and education using standard coding schemes, such as the International Standard Industrial Classification (ISIC), the International Standard Classification of Occupations (ISCO), and the International Standard Classification of Education (ISCED). Coding schemes were provided by Statistics Canada for all open-ended items, and countries were given specific instructions about the coding of such items.

In order to facilitate comparability in data analysis, each ALL country was required to map its national dataset into a highly structured, standardized record layout. In addition to specifying the position, format, and length of each field, the international record layout included a description of each variable and indicated the categories and codes to be provided for that variable. Upon receiving a country's file, Statistics Canada performed a series of range checks to ensure compliance to the prescribed format; flow and consistency edits were also run on the file. When anomalies were detected, countries were notified of the problem and were asked to submit cleaned files.

Scoring. Persons in each country charged with scoring received intense training, using the ALL scoring manual, in scoring responses to the open-ended items. They were also provided a tool for capturing closed format questions. To aid in maintaining scoring accuracy and comparability between countries, ALL introduced the use of an electronic bulletin board where countries could post their scoring questions and receive scoring decisions from the domain experts. This information could be seen by all countries, who could then adjust their scoring.
To further ensure quality, countries were monitored as to the quality of their scoring in two ways.

First, within a country, at least 20 percent of the tasks had to be rescored. Guidelines for intra-country rescoring involved rescoring a larger portion of booklets at the beginning of the scoring process to identify and rectify as many scoring problems as possible. In a second phase, countries selected a smaller portion of the next third of the scoring booklets; this phase was viewed as a quality monitoring measure and involved rescoring a smaller portion of booklets regularly to the end of the rescoring activities. The two sets of scores needed to match with at least 95 percent accuracy before the next step of processing could begin. In fact, most of the intra-country scoring reliabilities were above 95 percent. Where errors occurred, a country was required to go back to the booklets and rescore all the questions with problems and all the tasks that belonged to a problem scorer.

Second, an international rescore was performed. Each country had 10 percent of its sample rescored by scorers in another country. For example, a sample of task booklets from the United States was rescored by the persons who had scored Canadian English booklets, and vice versa. The main goal of the rescore was to verify that no country scored consistently differently from another country. Intercountry score reliabilities were calculated by Statistics Canada and the results were evaluated by the ETS. Again, strict accuracy was demanded: a 90 percent correspondence was required before the scores were deemed acceptable. Any problems detected had to be rescored.

Estimation Methods

Weighting was used in ALL to adjust for sampling and nonresponse. Responses to the literacy tasks were scored using item response theory (IRT) scaling. A multiple imputation procedure based on plausible values methodology was used to estimate the literacy proficiencies of individuals who completed literacy tasks.

Weighting. Each participating country in ALL used a multistage probability sample design with stratification and unequal probabilities of respondent selection. Furthermore, there was a need to compensate for the nonresponse that occurred at varying levels. Therefore, the estimation of population parameters and the associated standard errors was dependent on the survey weights. All participating countries used the same general procedure for calculating the survey weights. However, each country developed the survey weights according to its particular probability sample design. In general, two types of weights were calculated by each country: population weights that are required for the production of population estimates and jackknife replicate weights that are used to derive the corresponding standard errors.

Population weights. For each respondent record, the population weight was created first by calculating the theoretical or sample design weight, then by deriving a base sample weight by mathematically adjusting the theoretical weight for nonresponse. The base weight is the fundamental weight that can be used to produce population estimates. However, in order to ensure that the sample weights were consistent with a country's known population totals (i.e., benchmark totals) for key characteristics, the base sample weights were ratio-adjusted to the benchmark totals.

Jackknife weights. It was recommended that 10 to 30 jackknife replicate weights be developed for use in determining the standard errors of the survey estimates. Switzerland produced 15 jackknife replicate weights. The remaining countries produced 30 jackknife replicate weights.

Scaling. The results of ALL are reported along four scales—two literacy scales (prose and document), a single numeracy scale, and a scale capturing problem solving—with each ranging from 0 to 500 points. One might imagine these tasks arranged along their respective scale in terms of their difficulty for adults and the level of proficiency needed to respond correctly to each task. The procedure used in ALL to model these continua of difficulty and ability is IRT. IRT is a mathematical model used for estimating the probability that a particular person will respond correctly to a given task from a specified pool of tasks.

The scale value assigned to each item results from how representative samples of adults in participating countries perform on each item and is based on the theory that someone at a given point on the scale is equally proficient in all tasks at that point on the scale. For ALL, as for IALS, proficiency was determined to mean that someone at a particular point on the proficiency scale would have an 80 percent chance of answering items at that point correctly.

Just as adults within each participating country in ALL are sampled from the population of adults living in households, each task that was constructed and used in the assessment represents a type of task sampled from the domain or construct defined here. Hence, it is representative of a particular type of literacy, numeracy, or problem-solving task that is associated with adult contexts.

In an attempt to display the progression of complexity and difficulty from the lower end of each scale to the upper end, each proficiency scale was divided into levels. Both the literacy and numeracy scales used five levels, where Level 1 represents the lowest level of proficiency and Level 5 the highest. These levels are defined as follows: Level 1 (0 to 225), Level 2 (226 to 275), Level 3 (276 to 325), Level 4 (326 to 375), and Level 5 (376 to 500). The scale for problem solving used four levels, where Level 1 is the lowest level of proficiency and Level 4 the highest. These four levels are defined as follows: Level 1 (0 to 250), Level 2 (251 to 300), Level 3 (301 to 350), and Level 4 (351 to 500).

Since each level represents a progression of knowledge and skills, individuals within a particular level not only demonstrate the knowledge and skills associated with that level but the proficiencies associated with the lower levels as well. In practical terms, this means that individuals performing at 250 (the middle of Level 2 on one of the literacy or numeracy scales) are expected to be able to perform the average Level 1 and Level 2 tasks with a high degree of proficiency. A comparable point on the problem-solving scale would be 275. In ALL, as in IALS, a high degree of proficiency is defined in terms of a response probability of 80 percent. This means that individuals estimated to have a particular scale score are expected to perform tasks at that point on the scale correctly with an 80 percent probability. It also means they will have a greater than 80 percent chance of performing tasks that are lower on the scale. It does not mean, however, that individuals with given proficiencies can never succeed at tasks with higher difficulty values. It does suggest that the more difficult the task relative to their proficiency, the lower the likelihood of a correct response.

Imputation. A respondent had to complete the background questionnaire, correctly complete at least two out of six simple tasks from the core block of literacy tasks, and attempt at least five tasks per literacy scale in order for researchers to be able to estimate his or her literacy skills directly. Literacy proficiency data were imputed for individuals who failed or refused to perform the core literacy tasks and for those who passed the core block, but did not attempt at least five tasks per literacy scale. Because the model used to impute literacy estimates for nonrespondents relied on a full set of responses to the background questions, ALL countries were instructed to obtain at least a background questionnaire from sampled individuals. ALL countries were also given a detailed nonresponse classification to use in the survey.

Literacy proficiencies of respondents were estimated using a multiple imputation procedure based on plausible values methodology. Special procedures were used to impute missing cognitive data.

Literary proficiency estimation (plausible values). A multiple imputation procedure based on plausible

values methodology was used to estimate respondents' literacy proficiency in ALL. When a sampled individual decided to stop the assessment, the interviewer used a standardized nonresponse coding procedure to record the reason why the person was stopping. This information was used to classify nonrespondents into two groups: (1) those who stopped the assessment for literacy-related reasons (e.g., language difficulty, mental disability, or reading difficulty not related to a physical disability); and (2) those who stopped for reasons unrelated to literacy (e.g., physical disability or refusal). The reasons given most often by individuals for not completing the assessment were reasons related to their literacy skills; the other respondents gave no reason for stopping or gave reasons unrelated to their literacy.

When individuals cited a literacy-related reason for not completing the cognitive items, it implies that they were unable to respond to the items. On the other hand, citing reasons unrelated to literacy implies nothing about a person's literacy proficiency. Based on these interpretations, ALL adapted a procedure originally developed for the National Adult Literacy Survey to treat cases in which an individual responded to fewer than five items per literacy scale, as follows: (1) if the individual cited a literacy-related reason for not completing the assessment, then all consecutively missing responses at the end of the block of items were treated as wrong; and (2) if the individual cited reasons unrelated to literacy for not completing the assessment, then all consecutively missing responses at the end of a block were treated as "not reached."

Proficiency values were estimated based on respondents' answers to the background questions and the cognitive items. As an intermediate step, the functional relationship between these two sets of information was calculated, and this function was used to obtain unbiased proficiency estimates with reduced error variance. A respondent's proficiency was calculated from a posterior distribution that was the multiple of two functions: a conditional distribution of proficiency, given responses to the background questions; and a likelihood function of proficiency, given responses to the cognitive items.

Future Plans

The OECD plans to conduct another survey, the Program for the International Assessment for Adult Competencies (PIAAC). It is built on the knowledge and experiences gained from IALS and ALL. PIAAC will measure relationships between educational background, workplace experiences and skills, professional attainment, use of ICT, and cognitive skills in the areas of literacy, numeracy and problem-solving. The assessment will be administered to 5,000 adults from ages 16 to 65. Administration of the survey will occur in 2011, with results being released in early 2013.

5. DATA QUALITY AND COMPARABILITY

The literacy tasks contained in ALL and the adults asked to participate in the survey were samples drawn from their respective universes. As such, they were subject to some measurable degree of uncertainty. ALL implemented procedures to minimize both sampling and nonsampling errors. The ALL sampling design and weighting procedures assured that participants' responses could be generalized to the population of interest. Quality control activities were employed during interviewer training, data collection, and processing of the survey data.

Sampling Error

Because ALL employed probability sampling, the results were subject to sampling error. Although small, this error was higher in ALL than in most studies because the cost of surveying adults in their homes is so high. Most countries simply could not afford large sample sizes.

Each country provided a set of replicate weights for use in a jackknife variance estimation procedure.

Nonsampling Error

The key sources of nonsampling error in ALL were differential coverage across countries and nonresponse bias, which occurred when different groups of sampled individuals failed to participate in the survey. Other potential sources of nonsampling error included deviations from prescribed data collection procedures and errors of logic that resulted from mapping idiosyncratic national data into a rigid international format. Scoring error, associated with scoring open-ended tasks reliably within and between countries, also occurred. Finally, because ALL data were collected and processed independently by the various countries, the study was subject to uneven levels of commonplace data capture, data processing, and coding errors.

Coverage error. The design specifications for ALL stated that in each country the study should cover the civilian, noninstitutionalized population ages 16 to 65. It is the usual practice to exclude the institutionalized population from national surveys because of the difficulties in conducting interviews in institutional settings. Similarly, it is not uncommon to exclude

certain other parts of a country's population that pose difficult survey problems (e.g., persons living in sparsely populated areas). The intended coverage of the surveys generally conformed well to the design specifications: each of the ALL countries attained a high level of population coverage. However, it should be noted that actual coverage is generally lower than the intended coverage because of deficiencies in sampling frames and sampling frame construction (e.g., failures to list some households and some adults within listed households).

Nonresponse error. For ALL, several procedures were developed to reduce biases due to nonresponse, based on how much of the survey the respondent completed.

Unit nonresponse. The definition of a respondent for ALL was a person who partially or fully completed the background questionnaire. Unweighted response rates varied considerably from country to country, ranging from a high of 82 percent (Bermuda) to a low of 40 percent (Switzerland). The United States had an unweighted response rate of 66 percent (see table 17).

Several precautions were taken against nonresponse bias. Interviewers were specifically instructed to return several times to nonrespondent households in order to obtain as many responses as possible. In addition, all countries were asked to ensure that the address information provided to interviewers was as complete as possible in order to reduce potential household identification problems.

Item nonresponse. Not-reached responses were classified into two groups: nonparticipation immediately or shortly after the background information was collected; and premature withdrawal from the assessment after a few cognitive items were attempted. The first type of not-reached response varied a great deal across countries according to the frames from which the samples were selected. The second type of not-reached response was due to quitting the assessment early, resulting in incomplete cognitive data. Not-reached items were treated as if they provided no information about the respondent's proficiency, so they were not included in the calculation of likelihood functions for individual respondents. Therefore, not-reached responses had no direct impact on the proficiency estimation for subpopulations. The impact of not-reached responses on the proficiency distributions was mediated through the subpopulation weights.

Measurement error. Assessment tasks were selected to ensure that, among population subgroups, each literacy domain (prose, document, numeracy, and problem solving) was well covered in terms of difficulty, stimuli type, and content domain. The ALL item pool was developed collectively by participating countries. Items were subjected to a detailed expert analysis at ETS and vetted by participating countries to ensure that the items were culturally appropriate and broadly representative of the population being tested. For each country, experts who were fluent in both English and the language of the test reviewed the items and identified ones that had been improperly adapted. Countries were asked to correct problems detected during this review process. To ensure that all of the final survey items had a high probability of functioning well, and to familiarize participants with the unusual operational requirements involved in data collection, each country was required to conduct a pilot survey.

Table 17. Sample size and response rate for the United States for the Adult Literacy and Lifeskills Survey (ALL): 2003

Country	Population ages 16 to 65 (millions)	Initial sample size	Out-of-scope cases[1]	Number of respondents[2]	Unweighted response rate (percent)
United States	184	7,045	1,846	3,420	66

[1]Out-of-scope cases are those where the residents were not eligible for the survey, the dwelling could not be located, the dwelling was under construction, the dwelling was vacant or seasonal, or the cases were duplicates.

[2]A respondent's data are considered complete for the purposes of the scaling of a country's psychometric assessment data provided that at least the Background Questionnaire variables for age, gender, and education have been completed.

SOURCE: Desjardins, R., Murray, S., Clermont, Y., and Werquin, P. (2005). *Learning a Living: First Results of the Adult Literacy and Life Skills Survey*. Ottawa, Canada: Statistics Canada.

Although the pilot surveys were small and typically were not based strictly on probability samples, the information they generated enabled ETS to reject items, to suggest modifications to a few items, and to choose good items for the final assessment. ETS's analysis of the pilot survey data and recommendations for final test design were presented to and approved by participating countries.

6. CONTACT INFORMATION

For content information on ALL, contact:

Eugene Owen
Phone: (202) 502-7422
E-mail: eugene.owen@ed.gov

Mailing Address:

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

7. METHODOLOGY AND EVALUATION REPORTS

General

Desjardins, R., Murray, S., Clermont, Y., and Werquin, P. (2005). *Learning a Living: First Results of the Adult Literacy and Life Skills Survey*. Ottawa, Canada: Statistics Canada.

Lemke, M., Miller, D., Johnston, J., Krenzke, T., Alvarez-Rojas, L., Kastberg, D., and Jocelyn, L. (2005). *Highlights From the 2003 International Adult Literacy and Lifeskills Survey (ALL)-(Revised)* (NCES 2005-117rev). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Chapter 25: Progress in International Reading Literacy Study (PIRLS)

1. OVERVIEW

PROGRESS IN INTERNATIONAL READING AND LITERACY STUDY:

Three aspects of reading literacy:

- Purpose for reading
- Processes of comprehension
- Reading behaviors and attitudes

Four sets of questionnaires:

- Student questionnaire
- Learning to read (home) survey
- Teacher questionnaire
- School principal questionnaire

The Progress in International Reading Literacy Study (PIRLS) is a large international comparative study of the reading literacy of fourth-grade students. The study is conducted by the International Association for the Evaluation of Educational Achievement (IEA), with national sponsors in each participating jurisdiction. The National Center for Education Statistics (NCES), in the Institute of Education Sciences at the U.S. Department of Education, is responsible for the implementation of PIRLS in the United States. Reading literacy is one of the most important abilities that students acquire as they progress through their early school years. It is the foundation for learning across all subjects, it can be used for recreation and for personal growth, and it equips young children with the ability to participate fully in their communities and the larger society. Participants in PIRLS include both countries and subnational entities, both of which are referred to as "jurisdictions." PIRLS focuses on the achievement and reading experiences of children in grades equivalent to fourth grade in the United States. The study includes a written test of reading comprehension and a series of questionnaires focusing on the factors associated with the development of reading literacy. PIRLS was first administered in 2001 to students in 35 jurisdictions and was administered again in 2006 to students in 45 jurisdictions. The next PIRLS is scheduled for 2011.

Purpose

PIRLS is a carefully constructed reading assessment, consisting of a test of the reading literacy of fourth-grade students and questionnaires to collect information about fourth-grade students' reading literacy performance. PIRLS has four goals: (1) develop internationally valid instruments for measuring reading literacy suitable for establishing internationally comparable literacy levels in each of the participating jurisdictions; (2) describe on one international scale the literacy profiles of fourth-graders in school in each of the participating jurisdictions; (3) describe the reading habits of fourth-graders in each participating jurisdiction; and (4) identify the home, school, and societal factors associated with the literacy levels and reading habits of fourth-graders in school.

Components

PIRLS focuses on three aspects of reading literacy: purposes for reading; processes of comprehension; student reading behaviors and engagement. The first two form the basis of the written test of reading comprehension. The student background questionnaire addresses the third aspect.

In PIRLS, purpose for reading refers to the two types of reading that account for most of the reading young students do, both in and out of school: (1) reading for literary experience, and (2) reading to acquire and use information. In the assessment, narrative fiction is used to assess students' ability to read for literary experience, while a variety of informational texts are used to assess students ability to acquire and use information while reading. The PIRLS assessment contains an equal proportion of texts assessing each purpose. Processes of comprehension refer

to ways in which readers construct meaning from the text. There are four comprehension processes: focusing on and retrieving specific ideas; making inferences; interpreting and integrating ideas and information; and examining or evaluating text features.

Assessment. The PIRLS assessment instruments include stories and informational texts at the fourth-grade level collected internationally. Students are asked to engage in a full repertoire of reading skills and strategies, including retrieving and focusing on specific ideas, making simple and more complex inferences, and examining and evaluating text features. The passages are followed by constructed-response and multiple-choice format questions about the text.

In PIRLS 2001, reading passages were printed in some students' assessment booklets, while other students were given the *PIRLS Reader*, a short anthology of a variety of reading texts, in addition to an assessment booklet. Using different booklets allows PIRLS to report results from more assessment items than can fit in one booklet, without making the assessment longer. To provide good coverage of each skill domain, the test items developed required over 5 hours of testing time. However, testing time was kept to 80 minutes for each student by clustering items in 8 blocks distributed across the 10 booklets, (9 student test booklets and the *PIRLS Reader*). Each student completed only one of the booklets. As a consequence, no student received all items, but each item was answered by a representative sample of students.

PIRLS 2006's design was built on PIRLS 2001. To evaluate changes in achievement over time, in 2006 new measuring scales were created in addition to the scale for reading achievement overall. To accommodate these changes, the booklet design expanded to include additional test booklets, and the total assessment time increased. PIRLS 2006 included 10 blocks, consisting of a reading passage and its accompanying questions. Four of the PIRLS 2001 test blocks were kept secure and carried forward for measuring trends in 2006, the six remaining blocks were redesigned. The new materials were added to reflect the broad approaches established for 2001, while refreshing and expanding the range of texts and devising items that brought out the qualities of each passage. The item blocks were then distributed across 13 booklets (including *PIRLS Reader*, a full color, magazine-style booklet) and each student was administered one of the booklets.

Questionnaires. Background questionnaires in PIRLS are administered to collect information about students' home and school experiences in learning to read. By gathering information about children's experiences (together with reading achievement on the PIRLS test), it is possible to identify the factors or combinations of factors that relate to high reading literacy. PIRLS 2001 and PIRLS 2006 administered questionnaires to students, teachers, and school principals. In jurisdictions other than the United States, a parent questionnaire is also administered. Additionally, PIRLS 2006 included a newly constructed curriculum questionnaire that provided information about the national context.

Student questionnaire. Each student taking the PIRLS reading assessment completes the student questionnaire. The questionnaire asks about aspects of students' home and school experiences, including instructional experiences and reading for homework, self-perceptions about and attitudes toward reading, out-of-school reading habits, computer use, home literacy resources, and basic demographic information, such as parents' educational level, language spoken at home, and student reading activities.

Learning to read (home) survey. The learning to read survey is completed by the parents or primary caregivers of each student taking the PIRLS reading assessment. It addresses child/parent literacy interactions, home literacy resources, parents' reading habits and attitudes, home/school connections, and basic demographic and socioeconomic indicators. This assessment was not administered in the United States in 2001 and 2006.

Teacher questionnaire. The reading teacher of each fourth-grade class sampled for PIRLS completes a questionnaire designed to gather information about classroom contexts for developing reading literacy. This questionnaire asks teachers about characteristics of the class tested (such as size, reading levels of the students, and language abilities of the students). It also asks about instructional time, materials and activities for teaching reading and promoting the development of students' reading literacy, and the grouping of students for reading instruction. Questions about classroom resources, assessment practices, and home/school connections are also included. The questionnaire also asks teachers for their views on opportunities for professional development and collaboration with other teachers and for information about their education and training.

School questionnaire. The principal of each school sampled for PIRLS responds to the school questionnaire. The questionnaire asks principals about enrollment and other school characteristics (such as where the school is located, resources available in the

surrounding area, and indicators of the socioeconomic background of the student body), characteristics of reading education in the school, instructional time, school resources (such as the availability of instructional materials and staff), home/school connections, and the school climate.

Curriculum questionnaire. First used in PIRLS 2006, this questionnaire focused on the nature of the development and implementation of a nationally (or regionally) defined reading curriculum in primary schools within each participating country.

In all, PIRLS takes 1½ to 2 hours of each student's time, including the assessment and background questionnaire.

In addition, system level information was provided by each participating country and published in the PIRLS 2001 Encyclopedia (Mullis et al. 2002) and the PIRLS 2006 Encyclopedia (Kennedy et al. 2007). The encyclopedias provide a description for each participating country of the policies and practices that guide school organization and classroom reading instruction in the lower grades.

Periodicity

PIRLS is administered once every 5 years, near the end of the school year in each jurisdiction. PIRLS was conducted in 2001 and 2006, and will be administered in the United States and other participating jurisdictions again in 2011.

2. USES OF DATA

PIRLS will help educators and policymakers by answering questions such as the following:

- How well do fourth-grade students read?
- How do students in one jurisdiction compare with students in another jurisdiction?
- Do fourth-grade students value and enjoy reading?
- Internationally, how do the reading habits and attitudes of students vary?

3. KEY CONCEPTS

International desired population. This is the grade or age level that each jurisdiction should address in its sampling activities. The international desired population for PIRLS 2001 was defined as all students enrolled in the upper of the two adjacent grades that contain the largest proportion of 9-year-olds at the time of testing. For PIRLS 2006, the international desired population was defined as all students enrolled in the grade that represents 4 years of schooling, counting from the 1st year of the International Standard Classification of Education (ISCED) Level 1, providing that the mean age at the time of testing was at least 9.5 years. For most jurisdictions, the target grade was the fourth grade or its national equivalent.

National desired population. PIRLS expects all participating jurisdictions to define their national desired population to correspond as closely as possible to the definition of the international desired population. For example, for PIRLS 2001, if the fourth grade was the upper of the two adjacent grades containing the greatest proportion of 9-year-olds in a particular jurisdiction, then students enrolled in fourth grade were the national desired population for that jurisdiction. For PIRLS 2006, if the fourth grade of primary school was the grade that represents 4 years of schooling in a particular jurisdiction (counting from the 1st year of ISCED Level 1), then students enrolled in fourth grade were the national desired population for that jurisdiction.

Although jurisdictions are expected to include all students in the target grade in their definition of the population, sometimes they have to reduce their coverage. Using its national desired population as a basis, each participating jurisdiction has to define its population in operational terms for sampling purposes. Ideally, the national defined population should coincide with the national desired population, although in reality there may be some school types or regions that cannot be included; consequently, the national defined population is usually a very large subset of the national desired population.

National Research Coordinators (NRCs) and data collection contractor. Each participating jurisdiction appoints a national research coordinator to monitor national data collection and processing in accordance with international standards. NCES contracts with a data collection firm to draw the samples, work with school coordinators, assemble and print the test booklets, and pack and ship the necessary materials to the sampled schools. The contractor is also

responsible for working with school coordinators, translating the test instruments, assembling and printing the test booklets, and packing and shipping the necessary materials to the sampled schools. They are also responsible for arranging the return of the testing materials from the school to the national center, preparing for and implementing the constructed-response scoring, entering the results into data files, conducting on-site quality assurance observations for a 10 percent sample of schools, and preparing a report on survey activities.

Reading literacy. The ability to use printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential. This definition goes beyond simply decoding and comprehending text to include a broad range of information-processing skills that adults use in accomplishing the range of tasks associated with work, home, and community contexts. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers, and for enjoyment. In PIRLS, there is a distinction between reading for literary experience and reading to acquire and use information.

4. SURVEY DESIGN

Target Population

In IEA studies, the target population for all jurisdictions is known as the *international desired population*. The detailed definitions of international desired population for PIRLS 2001 and 2006 are provided in the section of Key Concepts. For both PIRLS 2001 and 2006, the international desired population corresponds to the fourth grade in most jurisdictions, including the United States. This population was chosen because it represents an important transition point in children's development as readers. In most jurisdictions, by the end of fourth grade, children are expected to have learned how to read, and are now reading to learn.

Sample Design

Using its national desired population as a basis, each participating jurisdiction has to define its population in operational terms for sampling purposes. PIRLS participants are expected to ensure that the national defined population includes at least 95 percent of the national desired population. Exclusions (which should be kept to a minimum) can occur at the school level, within the sampled schools, or at both levels. Because the national desired population is restricted to schools that contain the required grade, schools not containing the target grade are considered to be outside the scope of the sample—not part of the target population.

In each jurisdiction, representative samples of students are selected using a two-stage sampling design. In the first stage, at least 170 schools are selected using probability proportional to size (PPS) sampling. Jurisdictions can incorporate in their sampling design important reporting variables (for example, urbanicity or school type) as stratification variables. In the second stage, one or two fourth-grade classes are randomly sampled in each school. This results in a sample size of at least 3,750 students in each jurisdiction. Some jurisdictions opt to include more schools and classes, enabling additional analyses, which results in larger sample sizes. In 2006, PIRLS required that all student sample sizes should not be less than 4,000 students.

In the United States in 2001, a nationally representative sample of 3,760 fourth-grade students from 170 schools was selected. The schools were randomly selected first, and then one or two classrooms were randomly selected within each school. In the United States in 2006, a nationally representative sample of 5,190 fourth-grade students from 180 schools was selected. The schools were randomly selected first, and then one or two classrooms were randomly selected within each school.

First sampling stage. The sample selection method used for the first sampling stage in PIRLS makes use of a systematic PPS technique. In order to use this method, it is necessary to have some measure of size (MOS) of the sampling units. Ideally, this is the number of sampling elements within the units (e.g., the number of students in the school in the target grade). If this is unavailable, some other highly correlated measure, such as total school enrollment, is used. The schools in each explicit stratum are listed in order of the implicit stratification variables, together with the MOS for each school. Schools are further sorted by MOS within implicit stratification variables. The cumulative MOS is a measure of the size of the population of sampling elements; dividing it by the number of schools to be sampled gives the sampling interval.

The first school is sampled by choosing a random number in the range between 1 and the sampling interval. The school whose cumulative MOS contains the random number is the sampled school. By adding the sampling interval to that first random number, a second school is identified. This process of consistently adding the sampling interval to the previous selection number results in a PPS sample of the required size.

Very large jurisdictions have an opportunity to introduce a preliminary sampling stage before sampling schools. The Russian Federation and the United States avail themselves of this option. In these jurisdictions, the first step is to draw a sample of geographic regions using PPS sampling. Then a sample of schools is drawn from each sampled region. This design is used mostly as a cost reduction measure, where the construction of a comprehensive list of schools would have been either impossible or prohibitively expensive. Also, the additional sampling stage reduces the dispersion of the school sample, thereby potentially reducing travel costs. Sampling guidelines are put in place to ensure that an adequate number of units will be sampled from this preliminary stage.

Second sampling stage. The second sampling stage consists of selecting classrooms within sampled schools. As a rule, one classroom per school is sampled, although some participants opt to sample two classrooms. All classrooms are selected with equal probabilities for all jurisdictions. It is suggested that any classroom smaller than half the specified minimum cluster size be combined with another classroom from the same grade and school.

Trends in IEA's Reading Literacy Study. PIRLS jurisdictions that earlier participated in the 1991 IEA Reading Literacy Study had the option of undertaking the Trends in IEA's Reading Literacy Study, which measured trends in reading achievement using IEA's 1991 reading test and student questionnaire. Since the target population for the Trends in IEA's Reading Literacy Study is similar (but not identical) to the PIRLS target population, it is possible to use the PIRLS school sample as the basis for the trend study sample. Accordingly, the sampling plan for the Trends in IEA's Reading Literacy Study is simple: select every second school sampled for PIRLS, and from each of these, sample one additional classroom from the target grade. Since the sample of schools for the Trends in IEA's Reading Literacy Study is essentially a subsample of the PIRLS sample of schools, most of the required sampling tasks are carried out during the PIRLS school sampling.

Assessment Design

The PIRLS International Study Center is responsible for the design, development, and implementation of the study—including developing the instruments and survey procedures, ensuring quality in data collection, and analyzing and reporting the study results. The PIRLS Reading Development Group contributes to the framework and reading test. Committee members review various drafts of the framework and assessment blocks, and review and endorse the final reading test. The PIRLS Questionnaire Development Group, comprising representatives from nine countries, helps develop the PIRLS questionnaires (including writing items and reviewing drafts of all questionnaires).

Development of framework and questions. At the heart of the PIRLS assessment is the definition of reading literacy established by the Reading Development Group and refined by National Research Coordinators. The PIRLS definition of reading literacy builds on the definition used in the 1991 IEA study, but elaborates on that definition by making specific reference to reading by children.

In accordance with the framework, the passages in the reading test are authentic texts drawn from children's storybooks and informational sources. Submitted and reviewed by the PIRLS jurisdictions, the passages represent a range of types of literary and informational texts. The literary passages include realistic stories and traditional tales, while the informational texts include chronological and nonchronological articles, biographical articles, and informational leaflets.

Two item formats are used to assess children's reading literacy—multiple-choice and constructed-response. Each type of item is used to assess both reading purposes and all four reading processes.

Matrix sampling. PIRLS has ambitious goals for covering the domain of reading literacy. The Reading Development Group felt that at least eight passages and items (four for each reading purpose) were needed to provide a valid and reliable measure of reading achievement. Since it would not be possible to administer the entire test to any one student, PIRLS used a matrix sampling technique to distribute the assessment material among students, yet retain linkages necessary for scaling the achievement data.

In PIRLS 2001, assessment material was divided into 40-minute "blocks," each comprised of a passage (a story or article) and items representing at least 15 score points. There were eight such blocks, four for each reading purpose. The eight assessment blocks were distributed across 10 test booklets, and each student completed one booklet in an 80-minute testing session. Each booklet contained two blocks—two literary, two informational, or one of each—and most blocks appeared in three booklets. One of the 10 booklets was the *PIRLS Reader*, a color booklet containing two reading passages; the test items for it were located in a separate booklet. The two blocks for the *Reader* appeared only in that booklet. The distribution of blocks across booklets "links" the booklets to enable

the achievement data to be scaled using Item Response Theory (IRT) methods.

The new material developed for PIRLS 2006 was combined with the four secure blocks retained from the 2001 assessment, providing an overall assessment that would allow the calculation of trends over 5 years. The PIRLS 2006 reading assessment was comprised of 13 booklets, one of which was administered to each student. Each booklet contained two blocks, comprised of a story or article followed by a series of questions pertaining to the text passage. In 2006, there were 10 blocks in total (5 for each reading purpose), which were systematically rotated throughout the booklets. As in 2001, the two blocks for the *Reader* appeared only in that booklet.

Data Collection and Processing

Reference dates. PIRLS is administered near the end of the school year in each jurisdiction. For PIRLS 2001, in jurisdictions in the Northern Hemisphere (where the school year typically ends in May or June), the assessment was conducted in April, May, or June 2001.

In the PIRLS 2006, jurisdictions in the Northern Hemisphere conducted the assessment between March and May 2006. In the United States, data collection began slightly earlier and ended in early June. In the Southern Hemisphere, the school year typically ends in November or December; in these jurisdictions, the assessment was conducted in October or November in 2001 and in October and November in 2005.

Data collection. Each jurisdiction is responsible for carrying out all aspects of the data collection, using standardized procedures developed for the study. Manuals provide explicit instructions to the NRCs and their staff members on all aspects of the data collection—from contacting sampled schools to packing and shipping materials to the IEA Data Processing Center for processing and verification. Manuals are also prepared for test administrators and for individuals in the sampled schools who work with the national centers to arrange for the data collection within the schools. These manuals address all aspects of the assessment administration within schools (including test security, distribution of booklets, timing and conduct of the testing session, and returning materials to the national center).

The PIRLS International Study Center places great emphasis on monitoring the quality of the PIRLS data collection. In particular, the Study Center implements an international program of site visits, whereby international Quality Control Monitors (QCMs) visit a sample of 15 schools in each jurisdiction and observe the test administration. In addition to the international program, NRCs are also expected to organize an independent national quality control program based upon the international model. The latter program requires national QCMs to document data collection activities in their jurisdiction. The national QCMs visit a random sample of 10 percent of the schools (in addition to those visited by the international QCMs) and monitor the testing sessions—recording their observations for later analysis.

Editing. To ensure the availability of comparable, high-quality data for analysis, PIRLS takes rigorous quality control steps to create the international database. PIRLS prepares manuals and software for jurisdictions to use in creating and checking their data files, so that the information will be in a standardized international format before being forwarded to the IEA Data Processing Center (DPC) in Hamburg for creation of the international database. Upon arrival at the DPC, the data undergo an exhaustive cleaning process involving several iterative steps and procedures designed to identify, document, and correct deviations from the international instruments, file structures, and coding schemes. The process also emphasizes consistency of information within national datasets and appropriate linking among the student, parent, teacher, and school data files.

Throughout the process, the data are checked and double-checked by the IEA Data Processing Center, the International Study Center, and the national centers. The national centers are contacted regularly and given multiple opportunities to review the data for their jurisdictions. In conjunction with the IEA Data Processing Center, the International Study Center reviews item statistics for each cognitive item in each jurisdiction to identify poorly performing items. In general, the items exhibit very good psychometric properties in all jurisdictions.

Estimation Methods

Weighting. Sampling weights are calculated according to a three-step procedure involving selection probabilities for schools, classrooms, and students.

School weight. The first step consists of calculating a school weight, which also incorporates weighting factors from any additional front-end sampling stages, such as districts or regions. A school-level participation adjustment is then made to the school weight to compensate for any sampled schools that do not participate. This adjustment is calculated independently for each explicit stratum.

The PIRLS sample design requires that school selection probabilities be proportional to the school size, defined as enrollment in the target grade. For jurisdictions with a preliminary sampling stage (such as the United States and the Russian Federation), the basic first-stage weight also incorporates the probability of selection in this preliminary stage. The first-stage weight in such cases is simply the product of the "region" weight and the first-stage weight.

In some jurisdictions, schools are selected with equal probabilities. This generally occurs when a large sampling ratio is used. Also, in some jurisdictions, explicit or implicit strata are defined to deal with very large schools or small schools. Equal probability sampling is necessary in these strata.

First-stage weights are calculated for all sampled and replacement schools that participate. A school-level participation adjustment is required to compensate for those schools that are sampled but do not participate and, hence, are not replaced. Sampled schools that are found to be ineligible are removed from the calculation of this adjustment. The school-level participation adjustment is calculated separately for each explicit stratum.

Classroom weight. In the second step, a classroom weight reflecting the probability of the sampled classroom(s) being selected from all the classrooms in the school at the target grade level is calculated. All classrooms are sampled with equal probability. No classroom-level participation adjustment is necessary, since in most cases a single classroom is sampled in each school. If a school agrees to take part in the study, but the classroom refuses to participate, adjustment for nonparticipation is made at the school level. If one of two selected classrooms in a school does not participate, then the classroom weight is calculated as though a single classroom has been selected in the first place. The classroom weight is calculated independently for each school.

Student weight. Because intact classrooms are sampled in PIRLS, each student in the sampled classrooms is certain of selection, so the base student weight is 1.0. However, as a third and final step, a nonparticipation adjustment is made to compensate for students who do not take part in the testing. This is calculated independently for each sampled classroom. The basic sampling weight attached to each student record is the product of the three intermediate weights: the first-stage (school) weight, the second-stage (classroom) weight, and the third-stage (student) weight.

Overall sampling weight. The overall student sampling weight is the product of the three weights, including the nonparticipation adjustments.

Scaling. The primary approach to reporting PIRLS achievement data is based on IRT scaling methods. The IRT analysis provides a common scale on which performance can be compared across countries. Student reading achievement is summarized using a family of IRT models. In 2006 PIRLS, 2- and 3-parameter logistic IRT models were used for dichotomously scored items, and generalized partial credit models for constructed-response items with two or three available score points. The IRT methodology is preferred for developing comparable estimates of performance for all students, since students respond to different passages and items depending upon which of the test booklets they receive. This methodology produces a score by averaging the responses of each student to the items that he or she takes in a way that takes into account the difficulty and discriminating power of each item. The approach followed in PIRLS uses information from the background questionnaires to provide improved estimates of student performance (a process known as conditioning) and multiple imputation to generate student scores (or "plausible values") for analysis and reporting.

In addition to providing a basis for estimating mean achievement, scale scores permit estimates of how students within jurisdictions vary and provide information on percentiles of performance. Treating all participating jurisdictions equally, the PIRLS scale average across jurisdictions was set to 500 and the standard deviation to 100. Since the jurisdictions vary in size, each jurisdiction is weighted to contribute equally to the mean and standard deviation of the scale. The average and standard deviation of the scale scores are arbitrary and do not affect scale interpretation.

In the PIRLS 2001 analysis, achievement scales were produced for each of the two reading purposes—reading for literary experience and reading for information—as well as for reading overall. The PIRLS 2006 reading achievement scales were designed to provide reliable measures of student achievement common to both the 2001 and 2006 assessments, based on the metric established originally in 2001. In 2006 PIRLS, in addition to the scale for reading achievement overall, IRT scales were created to measure changes in achievement in the two purposes of reading and two overarching reading processes.

Imputation. No imputations are generated for missing values. However, multiple imputations are used to

generate student scores (or "plausible values") for analysis and reporting.

The PIRLS item pool is far too extensive to be administered in its entirety to any one student, and so a matrix-sampling test design was developed whereby each student is given a single test booklet containing only a part of the entire assessment. The results for all of the booklets are then aggregated using IRT techniques to provide results for the entire assessment. Since each student responds to a subset of the assessment items, multiple imputations (the generation of "plausible values") are used to derive reliable estimates of student performance on the assessment as a whole. Since every student proficiency estimate incorporates some uncertainty, PIRLS follows the customary procedure of generating five estimates for each student and using the variability among them as a measure of this imputation uncertainty, or error. In the PIRLS international reports (Mullis et al. 2003, 2007), the imputation error for each variable is combined with the sampling error for that variable to provide a standard error incorporating both.

5. DATA QUALITY AND COMPARABILITY

A group of distinguished international reading scholars, the Reading Development Group, was formed to construct the PIRLS framework and endorse the final reading assessment. Each jurisdiction followed internationally prescribed procedures to ensure valid translations and representative samples of students. The national QCMs compared the final version of the booklets with the international translation verifier's comments to ensure that their suggestions had been incorporated appropriately into the materials. The QCMs were then appointed in each jurisdiction to monitor the testing sessions at the schools to ensure that the high standards of the PIRLS data collection process were met.

Sampling Error

The standard errors of the reading proficiency statistics reported by PIRLS include both sampling and imputation variance components.

When, as in PIRLS, the sampling design involves multistage cluster sampling, there are several options for estimating sampling errors that avoid the assumption of simple random sampling. The jackknife repeated replication technique (JRR) is chosen by PIRLS because it is computationally straightforward and provides approximately unbiased estimates of the sampling errors of means, totals, and percentages.

The particular application of the JRR technique used in PIRLS is termed a paired selection model because it assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sample design, with each pair regarded as members of a pseudo-stratum for variance estimation purposes. When used in this way, the JRR technique appropriately accounts for the combined effect of the between- and within-PSU contributions to the sampling variance. The general use of JRR entails systematically assigning pairs of schools to sampling zones, and randomly selecting one of these schools to have its contribution doubled and the other to have its contribution zeroed, so as to construct a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for the original sample, and once again for each pseudo-replicate sample. The variation between the estimates for each of the replicate samples and the original sample estimate is the jackknife estimate of the sampling error of the statistic.

To apply the JRR technique used in PIRLS 2001 and PIRLS 2006, the sampled schools were paired and assigned to a series of groups known as "sampling zones." In total, 75 zones were used, allowing for 150 schools per jurisdiction. When more than 75 zones were constructed, they were collapsed to keep the total number to 75. For more information on sampling error, see the PIRLS technical reports (Martin, Mullis, and Kennedy 2003, 2007).

Imputation error. For each of the PIRLS reading scales, reading overall, and literary and informational reading, the IRT scaling procedure yields five imputed scores or plausible values for every student. The difference between the five values reflects the degree of uncertainty in the imputation process.

The general procedure for estimating the imputation variance using plausible values is the following. First compute the statistic (t) for each set of plausible values (M). The statistic t_m, where m = 1, 2, …, 5, can be anything estimable from the data, such as a mean, the difference between means, percentiles, and so forth. Once the statistics are computed, the imputation variance is then computed as

$$Var_{imp} = (1 + 1/M)Var(t_m)$$

where M is the number of plausible values used in the calculation, and $Var(t_m)$ is the variance of the estimates computed using each plausible value.

Nonsampling Error

Due to the particular situations of individual PIRLS jurisdictions, sampling and coverage practices have to be adaptable in order to ensure an internationally comparable population. As a result, nonsampling errors in PIRLS can be related both to coverage error and nonresponse.

Coverage error. PIRLS expects all participating jurisdictions to define their national desired population to correspond as closely as possible to its definition of the international desired population. Although jurisdictions are expected to include all students in the target grade in their definition of the population, sometimes they have to reduce their coverage. Although jurisdictions were expected to do everything possible to maximize coverage of the population by the sampling plan, schools could be excluded if they were in geographically remote regions, if they were of extremely small size, if they offered a curriculum or a school structure that was different from that found in the mainstream education system, or if they provided instruction only to students in the categories defined as "within-school exclusions."

Table 18. Weighted U.S. response rates for 2001 and 2006 PIRLS assessments

Year	School response rate	Student response rate	Overall response rate
2001	86	96	83
2006	86	95	82

NOTE: All weighted response rates refer to final adjusted weights. Response rates were calculated using the formula developed by the IEA for PIRLS. The standard NCES formula for computing response rates would result in a lower school response rate. Response rates are after replacement.
SOURCE: Martin, M.O., Mullis, I.V.S., and Kennedy, A.M. (Eds.). (2003). *PIRLS 2001 Technical Report*. Boston College, International Study Center. Chestnut Hill, MA. Baer, J., Baldi, S., Ayotte, K., and Green, P. (2007). *The Reading Literacy of U.S. Fourth-Grade Students in an International Context: Results From the 2001 and 2006 Progress in International Reading and Literacy Study (PIRLS)* (NCES 2008-017). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Within-school exclusions were limited to students who, because of some disability, were unable to take the PIRLS tests, including educable mentally disabled students, functionally disabled students, and non-native-language speakers.

Nonresponse error.

Unit nonresponse. Unit nonresponse error results from nonparticipation of schools and students. Weighted and unweighted school and student response rates for PIRLS are computed for each participating jurisdiction. To monitor school participation, three school participation rates are computed: one using only originally sampled schools; one using sampled and first replacement schools; and one using sampled and both first and second replacement schools. Student participation rates are also computed, as are overall participation rates.

The minimum acceptable school-level response rate, before the use of replacement schools, was set at 85 percent. Likewise, the minimum acceptable student-level response rate was set at 85 percent. Jurisdictions understood that the goal for sampling participation was 100 percent of all sampled schools and students. Guidelines for reporting achievement data for jurisdictions securing less than full participation were modeled after IEA's Trends in International Mathematics and Science Study (TIMSS). Jurisdictions were assigned to one of three categories on the basis of their sampling participation. Jurisdictions in Category 1 were considered to have met the PIRLS sampling requirements and to have an acceptable participation rate. Jurisdictions in Category 2 met the sampling requirements only after including replacement schools. Jurisdictions that failed to meet the participation requirements, even with the use of replacement schools, were assigned to Category 3.

In 2001, almost all jurisdictions met the PIRLS sampling requirements and belonged in Category 1. Because they met the sampling requirements only after including replacement schools, England, the Netherlands, and the United States belonged in Category 2. Although Morocco and Scotland had overall weighted participation rates of 69 and 74 percent, respectively (even after including replacement schools), it was decided that these rates did not warrant the placement of the jurisdictions in Category 3. Instead, the results for Morocco and Scotland were annotated to indicate that they nearly satisfied the guidelines for sample participation rates after including replacement schools.

In 2006, almost all jurisdictions met the PIRLS sampling requirements and belonged in Category 1. Because they met the sampling requirements only after including replacement schools, Scotland, the United States, the Netherlands, and Belgium (Flemish) were placed in Category 2. Although Norway had overall participation rates after including replacement schools of just below 75 percent (71 percent), it was

decideddduring the sampling adjudication that this rate did not warrant placement in Category 3. Instead, the results for Norway were annotated in the 2006 international report similarly to what was done for Morocco and Scotland in 2001.

Data Comparability

IEA Reading Literacy Study and PIRLS. In 1991, the IEA launched the Reading Literacy Study, which assessed the reading literacy of 4th- and 9th-grade students in 32 jurisdictions. In 2001, IEA launched PIRLS in 35 jurisdictions. Although built on the foundation of the 1991 study, PIRLS is a new and different study, with a new assessment framework describing the interaction between two major reading purposes (literary and informative) and a range of four comprehension processes, an innovative reading test, and newly developed questionnaires for parents, students, teachers, and school principals.

Because the PIRLS 2001 reading test differed in a number of respects from the 1991 test, it was not possible to link the results of the two studies directly together. However, since PIRLS 2001 was scheduled to collect data on fourth-grade students 10 years after the 1991 Reading Literacy Study, PIRLS jurisdictions that participated in 1991 were given the opportunity of measuring changes in reading literacy achievement over that period by re-administering the 1991 reading literacy test to primary and elementary school students as part of the PIRLS data collection. The resulting study is known as the Trends in IEA's Reading Literacy Study. In 2001, nine jurisdictions replicated the 1991 Reading Literacy Study: Greece, Hungary, Iceland, Italy, New Zealand, Singapore, Slovenia, Sweden, and the United States. Conducted at the third or fourth grades (the grade with the most 9-year-olds), the study assessed student reading in three major domains: narrative texts, expository texts, and documents. Students completed a brief questionnaire about their home and school literacy activities and instruction. For more information on the trend study, see *Trends in Children's Reading Literacy Achievement 1991–2001: IEA's Repeat in Nine Countries of the 1991 Reading Literacy Study* (Martin et al. 2003). No such trend study was administered in conjunction with the 2006 PIRLS.

The United States conducted a study to compare the two international studies in the aspects of reading literacy each assessed, the types of texts they used in the assessments, and the types and difficulty of the questions they used. Both differences and similarities were found. The definitions of reading literacy were very similar. The types of passages used were similar, but in actually choosing and categorizing passages, the Reading Literacy Study emphasized the types of texts, while PIRLS focused on purposes for reading. In most cases, the passages and texts in PIRLS were longer, more engaging, and more complex. The question taxonomies that were generated to frame the tasks in the assessments were very different. The Reading Literacy Study taxonomy had a text focus with activities such as verbatim responses, main theme, and locating information. The PIRLS taxonomy suggested more consideration of the readers' interaction with the passage, especially in the categories of "interpret and integrate ideas and information" and "examine and evaluate content, language, and textual elements." The use of a high number of constructed-response items permitted the PIRLS questions to tap a wider range of reading responses; this is supported by the limited analysis of a sample of questions using Wixso's Levels of Depth of Knowledge. In general, PIRLS called for a wider range of skills than did the Reading Literacy Study, especially skills requiring deeper thinking. Also, the PIRLS passages were presented in an engaging and authentic manner that might have improved students' motivation to read and respond to the texts. This is one area where the form of PIRLS might have contributed to students' level of performance. However, if students lacked the skills necessary to respond to the items, engaging texts would not have helped much. For more information on the comparison study, see the *PIRLS-IEA Reading Literacy Framework: Comparative Analysis of the 1991 IEA Reading Study and the Progress in International Reading Literacy Study* (Kapinus 2003).

National Assessment of Educational Progress (NAEP) and PIRLS. To date, there have been two studies undertaken to compare the frameworks, reading passages, and assessment items of NAEP and PIRLS. The first study compared NAEP 2002 and PIRLS 2001 at both the framework and item levels. The second study updates with analysis of the passages and item sets added in NAEP 2007 and PIRLS 2006.

Definitions and organizations. In terms of how the domain is defined, there is considerable overlap between the NAEP and PIRLS concepts of reading literacy. The differences are relatively minor: the PIRLS framework is more explicit about its targeting to young readers and acknowledges a more diverse set of reading contexts such as for personal enjoyment (versus the NAEP framework, which focuses more on school-based reading and is intended to be generally applicable across younger to older grades).

In terms of the organization of the frameworks, both NAEP and PIRLS are organized around two dimensional matrices, which specify processes (i.e., the

cognitive element) and the purposes or contexts for which students read. In particular, there are some notable differences at the framework level in how the processes (called aspects in NAEP) are broken out and elaborated. NAEP's four categories include: forming a general understanding, developing an interpretation, making reader-text connections, and examining content and structure. PIRLS' four categories include: locating and retrieving explicitly stated information, making straightforward inferences, interpreting and integrating ideas and information, and examining and evaluating content, language and textual elements. The key areas of difference are that there is no apparent counterpart in the NAEP framework to the PIRLS locate and retrieve category, and there is no explicit counterpart in the PIRLS framework to the NAEP category that requires readers to think beyond the text and apply it to the real world (i.e., make reader-text connections).

In terms of the purposes for which students read, both frameworks specify a literary purpose and an information-related purpose. While the literary purposes seem to be defined in a similar way across the assessments, the information-related purposes suggest slight differences. PIRLS assesses not just reading to acquire information, but also to use information, in a way that goes beyond NAEP's definition. At the older grades, the NAEP framework includes a "reading to perform a task" purpose, which focuses on reading to learn how to do something, which is more similar to the use information aspect of PIRLS' "reading to acquire and use information purpose.

Passage and item analyses. The types of passages included in NAEP and PIRLS reflect the purposes that are assessed. In NAEP, students are presented with short stories, legends, biographies, and folktales, as well as magazine articles that focus on people, places, and events of interest to children—to cover both its literary experience and information purposes. Similarly, PIRLS also presents narrative fiction, usually in the form of short stories, as well as informational articles and, distinct from NAEP, brochures to cover its two similar purposes. Both NAEP and PIRLS strive to be "authentic" in that they try to present passages and items that would be encountered in and out of school. NAEP specifically calls for the use of authentic texts, and all passages are shown as previously published and generally are not edited at all (in terms of content or formatting) for use in NAEP. PIRLS also strives to use previously published texts, but has a more liberal policy on editing and changing the format of the texts used—which is sometimes necessary in an international context in order to meet constraints of translation to multiple languages and for culturally diverse participants. U.S. experts who have examined the PIRLS passages have noted the more edited, and sometimes less continuous, nature of some of these than the NAEP passages, particularly among passages for information purpose.

Altogether, the NAEP and PIRLS fourth-grade assessments each include 10 reading passages, although each student receives only a subset of those passages. In terms of length, the PIRLS passages tend to be shorter than the NAEP passages, averaging 707 words per passage compared to NAEP's 823 words per passage. The PIRLS passages range from 403 to 855 words; NAEP passages range from 644 to 1,361 words.

Readability analyses also suggest that the PIRLS passages may be slightly easier than NAEP. On a very simple measure, for example, sentence counts show that the PIRLS passages, with a higher number of sentences per 100 word sample, consist of shorter sentences on average than do the NAEP passages. On other more elaborate measures, such as Fry and Flesch analyses, which use sentence count along with syllable count to determine a corresponding age and grade level for each text, PIRLS passages are calculated to be about one grade level below the NAEP passages. Finally, a Lexile measure, which indicates the reading demand of the text in terms of semantic difficulty (vocabulary) and syntactic complexity (sentence length) and which is more recently developed and normed than the other measures, also suggests that the PIRLS passages are suitable for one to two grades below those from NAEP. It should be noted, however, that both assessments do include a range of passages suited below and above the targeted grade level to capture the range of reading ability.

Each of these passages has items associated with it—approximately 12-13 per passage in PIRLS and 10 per passage in NAEP. The two assessments are similar in that the majority of items on both assessments require students to develop an interpretation about what they have read, although there is a greater emphasis on this in NAEP, with 69 percent of items classified as such compared to 60 percent of the PIRLS items. PIRLS also has a notably smaller percentage of items classified as forming a general understanding or making reader text connections, having half or less the percentage NAEP has in those categories. One of the major differences between the two assessments, however, is that there are a number of PIRLS items (21 percent) that do not fit on the NAEP framework at all. In nearly all cases, these are items that ask the reader to retrieve explicitly stated information, which is not a skill delineated in the NAEP framework or found in its items.

For more information on the similarities and differences between PIRLS and NAEP, see *A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments* (Binkley and Kelly 2003), and *Comparing PIRLS and PISA with NAEP in Reading, Mathematics, and Science* (Stephens, and Coleman, 2007).

6. CONTACT INFORMATION

For content information about PIRLS, contact:

Stephen Provasnik
Phone: (202) 502-7480
E-mail: stephen.provasnik@ed.gov

Mailing Address:

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

7. METHODOLOGY AND EVALUATION REPORTS

Most of the technical documentation for PIRLS is published by the International Study Center at Boston College. The U.S. Department of Education, National Center for Education Statistics, is the source of several additional references listed below; these publications are indicated by an NCES number.

General

Kennedy, A.M., Mullis, I.V.S., Martin, M.O., and Trong, K.L. (2007). *PIRLS 2006 Encyclopedia: A Guide to Reading Education in the Forty PIRLS 2006 Countries*. Boston College, International Study Center. Chestnut Hill, MA.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., and Kennedy, A.M. (2003). *Trends in Children's Reading Literacy Achievement 1991–2001: IEA's Repeat in Nine Countries of the 1991 Reading Literacy Study*. Boston College, International Study Center. Chestnut Hill, MA.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., and Kennedy, A.M. (Eds.). (2003). *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary Schools in 35 Countries*. Boston College, International Study Center. Chestnut Hill, MA.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., and Flaherty, C.L. (2002). *PIRLS 2001 Encyclopedia: A Reference Guide to Reading Education in the Countries Participating in IEA's Progress in International Reading Literacy Study (PIRLS).* Boston College, International Study Center. Chestnut Hill, MA.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., and Foy, P. (2007). *PIRLS 2006 International Report: IEA's Progress in International Reading Literacy Study in Primary Schools in 40 Countries*. Boston College, International Study Center. Chestnut Hill, MA.

Uses of Data

Foy, P., and Kennedy, A.M. (2008). *PIRLS 2006 User Guide for the International Database.* Boston College, International Study Center. Chestnut Hill, MA.

Gonzalez, E.J., and Kennedy, A.M. (2003). *PIRLS 2001 User Guide for the International Database.* Boston College, International Study Center. Chestnut Hill, MA.

Survey Design

Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O., and Sainsburg, M. (2001). *Framework and Specifications for PIRLS Assessment 2001—2nd Edition*. Boston College, International Study Center. Chestnut Hill, MA.

Martin, M.O., Mullis, I.V.S., and Kennedy, A.M. (Eds.). (2003). *PIRLS 2001 Technical Report.* Boston College, International Study Center. Chestnut Hill, MA.

Martin, M.O., Mullis, I.V.S., and Kennedy, A.M. (Eds.). (2007). *PIRLS 2006 Technical Report.* Boston College, International Study Center. Chestnut Hill, MA.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O., and Sainsburg, M. (2006). *PIRLS 2006 Assessment Frameworks, 2nd Edition.* Boston College, International Study Center. Chestnut Hill, MA.

Ogle, L.T., Sen, A., Pahlke, E., Jocelyn, L., Kastberg, D., Roey, S., and Williams, T. (2003). *International Comparisons in Fourth-Grade Reading Literacy: Findings From the Progress in International Reading Literacy Study of 2001* (NCES 2003-073). National Center for Education Statistics, Institute of

Education Sciences, U.S. Department of Education. Washington, DC.

Data Quality and Comparability

Baer, J., Baldi, S., Ayotte, K., and Green, P. (2007). *The Reading Literacy of U.S. Fourth-Grade Students in an International Context: Results From the 2001 and 2006 Progress in International Reading and Literacy Study (PIRLS)* (NCES 2008-017). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Binkley, M., and Kelly, D.L. (2003). *A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments* (NCES Working Paper 2003-10). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Kapinus, B. (2003). *PIRLS-IEA Reading Literacy Framework: Comparative Analysis of the 1991 IEA Reading Study and the Progress in International Reading Literacy Study* (NCES Working Paper 2003-05). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Piesse, A., and Rust, K. (2003). *U.S. 2001 U.S. PIRLS Nonresponse Bias Analysis* (NCES 2003-21). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Stephens, M., and Coleman, M. (2007). Comparing PIRLS and PISA with NAEP in Reading, Mathematics, and Science (Working Paper). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Available at:http://nces.ed.gov/Surveys/PISA/pdf/comppaper12082004.pdf.